Modelling the structure of complex networks

Herlau, Tue

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Modelling the structure of complex networks

Tue Herlau

Kongens Lyngby 2014
IMM-PhD-2014-339

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3351
compute@compute.dtu.dk
www.compute.dtu.dk IMM-PhD-2014-339

Summary (English)

A complex network is a set of distinct entities which interacts in a quantifiable manner. Representing systems as complex networks have become increasingly popular in a variety of scientific fields including biology, social sciences and economics. Complex networks have simultaneously been studied as mathematical objects of their own right and as a result, there has been both an increased demand for statistical methods for modelling complex networks as well as a quickly growing mathematical literature on the subject.

In this dissertation we explore aspects of modelling complex networks from a probabilistic perspective. The first two chapters will focus on the use of probabilistic methods for inference problems. We will consider a justification of probabilistic methods from the perspective of consistency and as a general method of updating beliefs. The next chapters will treat some of the various symmetries, representer theorems and probabilistic structures often used when modelling complex networks, the construction of sampling methods and various network models.

The introductory chapters will provide context for the included written work on the topics of (i) updating beliefs (ii) construction of samplers for partition-based problems (iii) applying non-parametric methods for modelling stationary and temporal network data.

Summary (Danish)

Et komplekst netværk er en samling af enheder som interagerer på en kvantificerbar måde. Det er stadigt mere populært at repræsentere systemer som komplekse netværk på tværs af en række videnskabelige discipliner herunder biologi, sociale videnskaber og økonomi. Parallelt med denne udvikling bliver komplekse netværk uafhængigt studeret som matematiske objekter. Derved er der både en stigende efterspørgsel efter statistiske metoder for modellering af komplekse netværk og en hastigt voksende matematisk litteratur om emnet.

I denne afhandling undersøger vi forskellige aspekter af modellering af komplekse netværk fra et statistisk perspektiv. De første to kapitler vil fokusere på at retfærdiggøre brugen af sandsynlighedsbaserede metoder til at drage konklusioner. Vi vil se på berettigelsen af sandsynlighedsbaserede metoder udfra et krav om konsistens og som en general metode til at opdatere tildelinger af sandsynligheder. De næste kapitler vil omhandle de forskellige symmetrier, repræsentationssætninger og sandsynlighedsteoretiske strukturer ofte brugt i modelleringen af komplekse netværk, konstruktion af sampling-baserede metoder og forskellige netværksmodeller.

De introducerende kapitler vil udgøre kontekst for det inkluderede skrevne arbejde som omhandler emnerne (i) at opdatere sandsynlighedstildelinger (ii) konstruktion af sampling-baserede metoder for partitions-baserede problemer (iii) at anvende ikke-parametriske metoder til modellering af stationær og temporal netværksdata.

List of Publications

Included work

The following manuscripts represents the main written work of this PhD project and were included in the version of the thesis presented at the defence. They will therefore be designated *"included work"* in the following.

[Herlau et al., 2015] Tue Herlau, Mikkel N. Schmidt, Morten Mørup, "Bayesian Dropout", *(in preperation)*, 2014.

[Herlau et al., 2014a] Tue Herlau, Morten Mørup, Yee Whye Teh, Mikkel N. Schmidt, "Adaptive Reconfiguration Moves for Efficient Markov Chain Sampling", *(in preperation)*, 2014.

[Herlau et al., 2012a] Tue Herlau, Morten Mørup, Mikkel N. Schmidt, Lars Kai Hansen, "Detecting Hierarchical Structure in Networks", *Cognitive Information Processing (CIP)*, 2012.

[Herlau et al., 2012b] Tue Herlau, Morten Mørup, Mikkel N. Schmidt, Lars K. Hansen, "Modelling Dense Relational Data", *Machine Learning and Signal Processing*, 2012.

[Herlau et al., 2013] Tue Herlau, Morten Mørup, Mikkel N. Schmidt, "Modeling Temporal Evolution and Multiscale Structure in Networks", *ICML* 2013.

[Herlau et al., 2014b] Tue Herlau, Mikkel N. Schmidt, Morten Mørup, "Infinite-degree-corrected stochastic block model", *Phys. Rev. E*, 2014.

Other manuscripts

[Schmidt et al., 2014] Schmidt, M. N., Herlau, T. and Mørup, M., "Probabilistic structural hierarchical clustering of normal relational data", *Cognitive Information Processing (CIP)*, 2014.

[Glückstad et al., 2014] Glückstad, F.K., Herlau, T., Schmidt, N. M., Mørup, M. "Cross-categorization of legal concepts across boundaries of legal systems: in consideration of inferential links". *Artificial Intelligence and Law, Springer, NY DOI: 10.1007/s10506-013-9150-2*, 2014.

[Glückstad et al., 2013c] Glückstad, F.K., Herlau, T., Schmidt, N. M., Rafal Rzepka, Kenji Araki, Mørup, M. "Analysis of conceptualization patterns across groups of people.", *Proceedings of 2013 Conference on Technologies and Applications of Artificial Intelligence (TAAI 2013), DOI 10.1109/.73 pp. 349-354*, 2013.

[Glückstad et al., 2013a] Glückstad, F.K., Herlau, T., Schmidt, N. M., Mørup, M., "Unsupervised Knowledge Structuring : Application of Infinite Relational Models to the FCA Visualization.", *The 9th International Conference on Signal Image Technology and Internet Based Systems.* SITIS 2013.

[Glückstad et al., 2013b] Glückstad, F.K., Herlau, T., Schmidt, N. M., Mørup, M., "Analysis of Subjective Conceptualizations Towards Collective Conceptual Modelling..", *Proceedings of the 27th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2013), Paper 795.* 2013.

[Ambrosen et al., 2013] Karen S. Ambrosen, Tue Herlau, Tim Dyrby, Mikkel N. Schmidt, Morten Mørup. "Comparing Structural Brain Connectivity by the Infinite Relational Model', *In Pattern Recognition in NeuroImaging (PRNI)*, 2013.

[Schmidt et al., 2012] Schmidt, M. N., Herlau, T. and Mørup, M., "Nonparametric Bayesian models of hierarchical structure in complex networks", *http://arxiv.org/pdf/1311.1033.pdf (Unpublished manuscript)*, 2012.

[Andersen et al., 2012] Andersen, K.W., Herlau, T. Mørup, M., Schmidt, M. N., Madsen, K. H., Lyksborg, M., Siebner, H., "Joint modelling of structural and functional brain networks", *NIPS workshop on Machine Learning and Interpretation in Neuroimaging*, 2012.

To my family,
to my Janina.

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling at the Technical University of Denmark in fulfilment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis consists of a brief treatment of selected topic relating to Bayesian methods and the modelling of complex networks and six research papers and manuscripts written during the period 2011-2014.

Lyngby, 31-March-2014

Tue Herlau

Acknowledgements

There are three people in particular to whom I wish to express my deepest gratitude. My supervisor and principal mentor Associated Professor Morten Mørup, for his encouragement and support in all matters. Morten has at all times stood ready with advice and words of comfort whenever my research needed a nudge in the right direction, and I am happy to have had the benefit of his knowledge and positive attitude. I also wish to thank Associated Professor Mikkel N. Schmidt for his continuous involvement in my project.

I only noticed recently both Morten and Mikkel have been involved in *all* my publications and have both played an indispensable role in everything I have accomplished during the last three years. Finally, I would like to thank my former supervisor Professor Lars Kai Hansen for creating an unique and pleasant environment as section leader. None of the following would have been possible or enjoyable without the help and support of these three people.

I would like to thank the members of the CogSys group, my three colleagues from room 321 Trine Abrahamsen, Bjarne Ørum Fruergaard and Toke Jansen Hansen for their friendship and my co-authors, Morten Mørup, Mikkel N. Schmidt, Lars Kai Hansen, Fumiko Glückstadt, Karen S. Ambrosen, Rafal Rzepka, Kenji Araki, Kasper W. Andersen, Kristoffer H. Madsen, Mark Lyksborg, Hartwig Siebner and Yee Whye Teh for the pleasure of their collaboration.

During my studies I spend six months at the department of statistics at the University of Oxford. I would sincerely like to thank Professor Yee Whye Teh for his invitation, hospitality and kind advice as well as all other members of the department of statistics for making me feel welcome and at home. I am also

grateful for the support of the Otto Mønsted Fond.

Contents

Chapter 1

Introduction

Complex networks denote systems composed of a set of entities (the vertices) that are interacting or related in a quantifiable manner (the edges). In the past decades there has been an increasing interest in describing real-world systems as networks and this has naturally generated a great interest in complex networks within many scientific disciplines. The use of complex networks to describe different systems has in turn led to the same network-related problems arising in different contexts. Examples of such problems includes (i) determine what constitutes important and general structural information in complex networks (ii) the study of complex networks as dynamic phenomena (iii) statistical concerns such as predicting missing information such as unobserved edges.

I do not know which of these questions are the more important. One of the great eye-opening experiences during my PhD studies was participating at the NETSCI 2013 conference in Copenhagen and realizing how many important questions, concerns, methods and results of the wider community I was unaware of.

In this work I will consider probabilistic modelling of networks from a machine-learning perspective. The two problems which has received the most attention from the machine learning perspective is to predict unobserved data as well as infer latent (often descriptive) structure from network data, for instance community structure. My work is focused on models which makes strong descriptive

structural assumptions often at the expense of better edge prediction.

A more fundamental problem one should invariantly consider is *why* one should use probabilistic models and not some other means to obtain the same goals. It was the arguments by Cox [1946] (as expressed by Jaynes [2003]) on the relationship between beliefs and probabilities that originally prompted me to change my area of study from physics to machine learning and this approach to machine learning, as something more fundamental than the engineering or mathematical challenges, has since been a recurrent interest and the first two chapters are dedicated to this subject.

1.1 Outline

My aims with this thesis are threefold. Firstly, to introduce relevant literature and relevant technical background for the included written work. Secondly, attempt to give a condensed account of some of the results and ways of thinking within the field which I believe to be relevant for a person who is either beginning a similar project as mine or has a general interest in the problems and results of a probabilistic approach to network science. Thirdly, that the thesis should be interesting to *write*. All chapters include results of great beauty from a wide range of fields and trying to give an account of these, however partial, have been a surprisingly rewarding experience. The downside of this way of selecting material is the thesis contains far too much for even an experienced writer to cover within its span. I have only attempted given the broadest overview and the thesis contains only two proofs which I absolutely could not make myself omit; the derivation of the rules of probability theory by Cox [1946] in chapter 2 and the derivation of Kullback-Leibner divergence as the unique belief-ranking functional by Shore and Johnson [1980] in chapter 3. Where this lack of self-restraint is particularly problematic is in the discussion of invariance in probability theory in chapter 4 and the reader should be aware the section cannot be used as a reference for an accurate statement of these results.

I have chosen to include some informal discussion and opinion throughout the thesis, especially in the introduction of chapter 2 and chapter 7. This have been done because I feel reflections on the nature of probabilistic methods in particular or machine learning in general is underpraised compared to the purely theoretical or engineering related concerns. At the very least, including some thoughts on the subject may induce someone to point out how I am wrong or unoriginal which would be worthwhile in itself. I hope the inclusion of this discussion is not too distracting from the main text.

The outline of the thesis is as follows. In chapter 2 and chapter 3 I will introduce probabilistic methods as tools for manipulating beliefs. The second part of the thesis, chapter 4, chapter 5 and chapter 6 is concerned with introducing tools from probability theory and applying them to network modelling. Finally chapter 7 contains a discussion and conclusion. To summarize the chapters are as follows:

Chapter 2, Beliefs, after some remarks on the nature of machine learning, this chapter is concerned with introducing probability theory as a consistency requirement for the (degree of) belief in propositions. The derivation is the axiomatic approach of Jaynes [2003]. The chapter conclude with examples of the use of symmetries to assign probabilities, the use of Bayesian methods for real-world problems and a connection of probability theory to machine learning exemplified by a basic network model.

Chapter 3, Assigning Beliefs, is concerned with arguing probability theory, as a consistency requirement on the assignment of beliefs, should be thought separate from the process of which beliefs are arrived at. After a brief discussion of this point I provide an account of the axiomatic approach to updating beliefs by Shore and Johnson [1980]. I include an argument for the central result which in the main follow that of Caticha and Giffin [2006] with some modifications. The second half of the chapter is concerned with applying the derived method to dropout [Hinton, Srivastava, Krizhevsky, Sutskever, and Salakhutdinov, 2012] and gives this method a probabilistic interpretation (see the included work in [Herlau et al., 2015]).

Chapter 4, Symmetries and invariance, discuss the concept of invariances and how they give rise to important representer theorems and their practical application for selected discrete probabilistic structures. I will briefly discuss important results including the De Finetti theorem and the Aldous-Hoover theorem, as well as introduce some basic results from the theory of exchangeable partitions and fragmentation which will be used later. The discussion will be informal and omits proofs.

Chapter 5, Inference, treats inference in probabilistic models exclusively from the perspective of Monte-Carlo sampling. After presenting important standard convergence results, I will very briefly introduce the concept of adaptive Markov-chain Monte Carlo sampling. The second part of the chapter is solely concerned with the problem of sampling partitions, in particular I will discuss a few of my own unsuccessful attempts that led to the proposed method, Adaptive Reconfiguration Moves [Herlau et al., 2014a].

Chapter 6, Networks, is concerned with a brief review of various network models proposed in recent years. The literature review is not intended to be complete, but rather to cover those models that are based on modelling assumptions which admits a representation compatible with the Aldous-Hoover theorem. The second part of the section will briefly discuss the far less well-explored topic of temporal network modelling. The section will briefly mention my own work on stationary relational network models [Herlau et al., 2012b,a, 2014b] as well as work on temporal hierarchical network modelling [Herlau et al., 2013].

Chapter 7, Discussion and Conclusion, as suggested by the name, this chapter contains closing remarks and an informal discussion of the previous chapters.

I will change to third-person pronouns for chapter 3–6 as these are concerned with either more general background information or work of which my co-authors share a significant part. I will however change back to first-person for the discussion and conclusion.

1.2 Included work

Included work refers to the main written work produced during my PhD which was included as appendices in the version of the thesis presented at my defence. I have deliberately chosen to let my own work play a very minor role in the following chapters since giving it a longer treatment at the expense of the other beautiful results would make this thesis far less enjoyable to read or write. The one exception to this rule is the paper *Adaptive Reconfiguration Moves for Efficient Markov Chain Sampling* [Herlau et al., 2014a] where I have included an informal discussion of my other less successful attempts which may benefit others who are working on the same problem. I will however include a brief synopsis of the included work below for easy reference:

[Herlau et al., 2015], Bayesian Dropout argues that probability theory should be seen as a consistency requirement and not as a learning method in and by itself. We argue, similar to Shore and Johnson [1980], that the process of learning is to update beliefs and this can and should be treated as an isolated problem. The derived method is used to derive a probabilistic variant of dropout, dubbed Bayesian Dropout, which is applied to linear and Logistic regression. In addition we discuss approaches for inference for the Bayesian Dropout including exact computation, analytical approximations and stochastic variational Bayes.

[Herlau et al., 2014a], Adaptive Reconfiguration Moves for Efficient Markov Chain Sampling discuss an application of adaptive Markov chain Monte Carlo for partition-based problems with special focus on the infinite relational model. The method discussed is based on evaluating multiple chains in parallel and use the similarity and dissimilarity between chains to construct transition kernels. The method is related to the split-merge sampler of Jain and Neal [2004] and the performance of the proposed method is evaluated on the infinite relational model and the Bernouilli mixture model.

[Herlau et al., 2012a], Detecting Hierarchical Structure in Networks in which we introduce model for networks where the vertices are partitioned into blocks, and the blocks of the partition are organized in a hierarchy which induces hierarchical structure in the network. By considering multifurcating hierarchies, the model is able to interpolate between the infinite relational model and past work on hierarchical modelling such as the approach by Clauset, Moore, and Newman [2008], the later model being limited to less expressive binary hierarchies. In addition, the use of hierarchies of communities allows more efficient sampling compared to the use of a hierarchy of vertices. On network data the use of multifurcating hierarchies is shown to give performance gains.

[Herlau et al., 2012b], Modelling dense relational data in this paper, we introduce a simple extension of the infinite relational model to continuous network data through the use of a Normal distribution for the observed data and a Normal-Wishart prior for the parameters of the normal distributions. We discuss the models relationship to a kernelized version of K-means and the resulting model is applied to continuous-valued relational data. This model was also used in other work [Glückstad, Herlau, Schmidt, and Morup, 2013a, Glückstad, Herlau, Schmidt, Mørup, Rzepka, and Araki, 2013c, Glückstad, Herlau, Schmidt, and Mørup, 2014].

[Herlau et al., 2013], Modeling temporal evolution and multiscale structure in networks in this work, we explore the application of hierarchies to temporal network data. In particular we propose a prior for temporally correlated hierarchies and discuss its statistical properties. This is to our knowledge the first work on temporal hierarchies from a Bayesian perspective. We build an efficient sampling method and apply the model to three larger temporal network datasets.

[Herlau et al., 2014b], Infinite-degree-corrected stochastic block model propose an extension to the infinite relational model in which, in addition to latent community structure, the degree of each vertex is modelled using vertex-specific parameters. By choosing a particular parametrization it is possible to analytically integrate out all infinite-dimensional latent

parameters such that only the community structure need to be sampled, allowing for a simple implementation.

Chapter 2

Beliefs

Few things has been as ingrained in popular thought as the distinction between mind and matter. While most thinkers have been willing to admit they do not know the fundamental reality of mind or matter, most have remained convinced there are important distinctions between the two and that both actually exists, and the problem of giving these two terms a definite meaning and determining their relationship has been at the forefront of philosophical and, when it was later invented, scientific inquiry for more than two and a half millennia [Livingston, 2004].

Scientific advances, in particular during the 20th century, has only made this problem more intriguing and is today known as the *hard problem of consciousness* [Chalmers, 1995]. Today our scientific knowledge of matter has increased to the point there is no experiment on earth whose outcome is not described by known theories[1], and the confirmation of the existence of the Higgs field at CERN in 2013 is only the latest of a series of remarkable predictions to come true. It goes without saying none of these experiments has encountered any "mind" (thoughts, intentions, consciousness, etc.) as distinct from, but influencing, matter and simply considering the precision of current experiments it is nearly impossible to imagine how any such thing could have remained undiscovered.

[1] Space, however, is another matter.

However, at the same time it is very difficult to see how known physics could give rise to a conscious observing mind. Indeed, if this was not our own immediate experience we would very likely conclude physics and biology distinctly *rules out* that possibility and while we can observe that our brains consists of neurons whose firing pattern correlates with behaviour there is of yet no theory of how the brain give rise to our thinking and experience of *being*.

In the absence of any general theoretical framework for how the mind works machine learning takes a pragmatic approach: We observe that humans are able to do certain things such as tell the number from licensing plates, we think of situations where this ability is useful and so we try to create a computer program that does the same. We observe humans drive cars and so we try to make computer programs which drives cars too. We see humans play Chess and so we try to make computer programs which play chess, or rather, we make computer programs that are good at archiving certain types of board positions. In all these circumstances the problem is posed as trying to map certain input to certain types of output. Sometimes the output is hard to define and sometimes the mapping is difficult to perform, however the point is the task can be described without any reference to a "mind".

However, even if we consider physics as having ruled out any satisfactory theory of the mind which *rests* on mental language such as desires, sensations, intentions, knowledge and so on having a causal influence over what we *do*, it does not follow there is no reason to suppose this same language (assuming it can be given definite meaning) cannot or should not have a legitimate place for describing cognition. As an analogy we can consider evolution. Today we have a fairly accurate account of the evolutionary process as it operates on life today from cellular chemistry to the species level. This was not always the case. Gregory Mendel, for instance, developed a primitive effective theory of genetics in 1865 by concluding there must be *something* like alleles of genes carrying inheritable traits in a certain manner Bowler [1989]. Certainly the concept of a gene was eventually tethered to a chemical foundation, however in terms of understanding evolution the *understanding* is exactly arrived at by being able to discuss evolution in terms of higher-level ideas such as alleles and genes which need no reference to the chemical foundation.

As an example of a mental term we can consider *truth*, in particular the relationship between the truth of various propositions. Deducing the truth of propositions from premises is undoubtedly a useful mental task, and logic is the formal analysis thereof. Obviously logic is not the *actual* cause of anything the brain does, however it remains a highly successful way to describe what rational thinking is in certain situations, in fact so successful that it attains a normative effect: If we encounter a situation which is amendable to logical analysis, logic tells us what we *ought* to believe, and if we nevertheless believe something

else, we will conclude we have made a mistake (been illogical). A problem one invariantly encounters when applying logic to every-day thinking is that we do not have perfect knowledge. For instance propositions like: *"It will rain tomorrow"* or *"my next hand will be a royal flush"* cannot *in principle* be known to be true or false with certainty. Commonly we would say we have a *belief* that, for instance, it will rain tomorrow. Analyzing the term belief (or degrees of belief) will be the subject of the remainder of this chapter. Our treatment will follow closely that of Cox [1946] (in particular as expressed in Jaynes [2003]) where the aim is to a set of *desiderata* formulated in natural language a notion of degree of belief *should* satisfy and then use these to provide a quantifiable characterization of *the degree of belief* of a proposition.

It is important to mention the approach of Cox is not the only way to define probabilities and it is not above philosophical or mathematical controversies. We will return to these issues in section 2.3.

2.1 Beliefs and probabilities

Propositions that we cannot know to be true or false with certainty plays a crucial role in every day reasoning. Consider the following example

> *A husband and wife agree the husband should not have more than two beers at the yearly Christmas party. The next day the wife finds the husband sleeping on the couch. There is a smell of day-old alcohol and cigarette smoke around him, the car is missing from the driveway and there is a red smear of what appears to be lipstick on his collar.*

We imagine his wife now believes he had more than two beers to drink. Later the same day we might consider the husband objecting to her beliefs. He might point out they are not the result of a logical deduction: For instance it could be the case a secretary had too much to drink at the Christmas lunch and slipped. As he heroically caught her she accidently spilled her drink over him and he got lipstick on his collar. It was then decided he, as the only sober person at the party, should drive the secretary and her husband home however after driving them home he got awfully sick and accidently locked his car keys and cell phone in his car and had to walk many miles to get a taxi. When he returned home much later he had simply thought it more considerate to sleep on the couch.

This explanation is *possible* however it is unlikely to be very convincing. To reason about what happened requires a system for reasoning under uncertainty

which extends classical propositional logic to the case where propositions cannot be known to be true or false. A key design crossroad revolve around the meaning of uncertainty. In this chapter we will be concerned with *graded states of belief* (i.e. plausibility, confidence, creditability) as opposed to *graded states of truth*. Multi-valued logic (also known as fuzzy logic) deals with the later [Zadeh, 1973, 1965, Hájek, 1998], see also *possibility theory* [Dubois and Prade, 1988] for a third notion of uncertainty. Multi-valued logic treats non-boolean propositions whose satisfaction is a matter of degree, whereas a theory of grades belief treats propositions where the uncertainty is induced by incomplete states of information. As a crude illustration of this distinction consider the example with the husband and suppose the wife considers the two propositions:

$$A: \quad \textit{"My husband had more than five beers"}$$
$$\mathcal{A}: \quad \textit{"My husband drank a lot"}.$$

In both cases she may not be certain if the proposition is true or false, however in the first case A this is due to imperfect information (if she had been able to follow her husband during the night *she would be certain* one way or the other) whereas for the second proposition $\mathcal{A}$, even if she knew exactly how many beers he had to drink (for instance five beers) she might still not be certain if this was really *"a lot"*. Put differently, if he had seven beers she would undoubtedly be more confident *that* would qualify as *a lot*. The accepted view today is both notions of uncertainty has an important role to play and deserve analysis, however the literature reflects a multitude of views on their relationship, foundations and interpretations [Zadeh, 1995, Hájek, 1998]. Loosely speaking, the study of graded truth has led to a greater variety of systems which have been less successfully applied in machine learning, however we will not attempt to survey the literature here.

Rather we will restrict ourselves to statements of the former kind A which are either true or false, and our uncertainty reflect a lack of knowledge [Jaynes, 2003, chapter 2]. Additional examples of such statements are:

$$B: \textit{"It will rain tomorrow"}$$
$$C: \textit{"If the procedure is administered to the patient she will test negative in two weeks"}$$
$$D: \textit{"The 918th decimal of } \pi \textit{ is 8"}$$

As indicated we will denote propositions with upper-case Latin letters, and we will assume they can be combined to form other propositions using the rules of classical propositional logic. Accordingly if A and B are propositions so is the conjunction AB (A *and* B), disjunction $A + B$ (A *or* B) and negation $\overline{A}$ (*not* A).

A notion of belief must encompass the notion propositions can be believed more or less strongly. For instance, if we have a particular belief there are 200 billion stars in our galaxy, we should have a lower degree of belief there are 400 billion. In other words, this example suggests beliefs come in degrees (they will be said to be *graded*) and that the degree can be compared.

It might be reasonable to wonder if we should be able to compare e.g. our degree of belief in whether it will rain tomorrow to our degree of belief there is life on Mars, however for simplicity we will assume all of our beliefs admits such a comparison. The simplest way to represent the *degree of* something is by a real number and thus we arrive at desiderate (I):

$$\text{(I)} \quad \textit{The belief of a proposition is represented as a single real number.} \tag{2.1}$$

All desiderata in this chapter is taken nearly verbatim from Jaynes [2003]. It should be stressed this assumption is not without controversy and likely the most important design choice. Firstly, as mentioned above it implies *universal comparability*, that is the degree of belief of any two propositions can be compared. See Fine [1973] for approaches which do not assume universal comparability. Secondly, it assumes the degree of belief is one-dimensional which for instance is not the case for Dempster-Shafer theory [Dempster, 1967, Shafer et al., 1976].

2.1.1 Relationship of beliefs

A second important property of beliefs is that they must relate to each other. For instance, if the wife in the example of the drunken husband holds a belief on whether her husband got drunk she must hold a belief whether her husband did not get drunk, and if she holds a belief in whether he smells like alcohol this must be related to her beliefs whether he had more than two beers the night before. In logic this relationship is that of proof [2]; given that certain propositions are taken as true, the truth (possibly the *degree of truth* in a multivalued logic) of other propositions is arrived at by proof [Hájek, 1998], however as beliefs express states of subjective knowledge we should not assume these are related through proof as the example with the husband and wife illustrates. Rather, following Cox [1946], Jaynes [2003] we will simply introduce a semantic for a proposition A given a proposition B is assumed to be true:

$$A|B \tag{2.2}$$

Read as *A given B is assumed true.* We will assume the type of propositions on which we hold beliefs all have the form of eq. (2.2). Since beliefs are represented

[2] i.e. a complete logic such as classical propositional logic [Hájek, 1998]

as real numbers according to desiderata (I) we will introduce the notation $(\cdot)$ for the numerical value of the belief, i.e.

$$(A|B) \tag{2.3}$$

denoted the *conditional belief* of A given B or *the degree of belief in A provided that B is true.* This type of entailment should also agree with common sense. If for instance A can be deduced from B, then if we have a high degree of belief in B clearly we should have a high degree of belief in A. We will capture this type of intuition in the somewhat informally stated desiderata

$$\text{(II)} \quad \textit{Beliefs are related to each other and the relationships between beliefs must qualitatively agree with commonsense.} \tag{2.4}$$

Evidently a great many things may be assumed under such a vague formulation and it is more important how the desiderata is invoked than how it is formulated above.

Notice we are not assuming a causal or logical relationship between the objects of A and B. To refer back to the example of the drunken husband an example of a non-causal relationship of $A|B$ may be:

$$A: \quad \textit{Husband was drunk last night} \tag{2.5}$$

$$B: \quad \textit{Lipstick on collar} \text{ AND } \textit{husband sleeping on couch} \text{ AND } \textit{husband smell of alcohol.} \tag{2.6}$$

Finally, if we consider beliefs to be related to each other, we will assume their relationship must be consistent. If there is more than one way to analyse a particular problem the result of the analysis must be consistent. Following Jaynes [2003] we assume the following three meanings of consistency:

$$\text{(IIIa)} \quad \textit{If conclusions can be arrived at in more than one way then all possible ways must lead to the same result.} \tag{2.7}$$

$$\text{(IIIb)} \quad \textit{One must admit all relevant information. If some types of information is relevant in simpler cases, the method of reasoning must consistently admit the same type of information in other situations.} \tag{2.8}$$

$$\text{(IIIc)} \quad \textit{Equivalent states of knowledge must be assigned equivalent states of belief. That is, if two problems is the same save labelling we must assign equivalent states of belief to both.} \tag{2.9}$$

2.1.2 The product rule

For beliefs to be useful[3] it should be possible to relate some beliefs to other beliefs. Consider the case where we wish to relate our belief in $AB|C$, or put in words, A and B given C, to other beliefs. In classical logic the deduction rule is that of the syllogism:

$$\begin{array}{l} B \text{ is true} \\ B \text{ implies } A. \\ \hline A \text{ is true.} \qquad \therefore \end{array} \tag{2.10}$$

This suggests one way to evaluate $AB|C$ is to first evaluate our degree of belief B is true given C, $(B|C)$. Then if B is true the only fact relevant to determine if AB is true is our belief in the truth of A; since we have assumed B is true this is $(A|BC)$.

This suggests the relevant degrees of belief are $(B|C)$ and $(A|BC)$. That is, we consider a relationship between the states of beliefs through a function F as

$$(AB|C) = F\big[(B|C), (A|BC)\big] \tag{2.11}$$

This expression may not have sufficient information on the right-hand side, however this will hopefully be apparent through later inconsistencies. On the other hand we might consider if we have included to much information and we could (for instance) have made do with a function only of the beliefs $(A|C)$ and $(B|C)$, however such an options may be ruled out by considering particular situations. Consider for instance the case of a flip of a coin where we consider the following propositions

$$\begin{aligned} A &: \textit{The side up is heads}, \\ B_1 &: \textit{The side down is heads}, \\ B_2 &: \textit{The side down is tails}. \end{aligned}$$

and C simply consist of our general knowledge about coins, flips etc. In this situation it is natural to consider the beliefs $(A|C), (B_1|C), (B_2|C)$ to be equivalent, however certainly $(AB_1|C)$ and $(AB_2|C)$ are not and so we need more information. A fuller treatment of various possibilities for the arguments of F is given by Tribus [1969].

Returning to eq. (2.11), by symmetry of A and B if eq. (2.11) holds so must

$$(AB|C) = F\big[(A|C), (B|AC)\big]. \tag{2.12}$$

[3]This section, and the remaining the derivation, is derived from Jaynes [2003]. See also Cox [1946].

Next consider three propositions A, B, C, D and our belief $(ABC|D)$. If we first consider BC as a single proposition and apply eq. (2.11) twice and next consider AB a single proposition and apply eq. (2.11) twice we get

$$(ABC|D) = F\big[(BC|D), (A|BCD)\big] = F\big\{F\big[(C|D), (B|CD)\big], (A|BCD)\big\} \quad (2.13a)$$
$$(ABC|D) = F\big[(C|D), (AB|CD)\big] = F\big\{(C|D), F\big[(B|CD), (A|BCD)\big]\big\} \quad (2.13b)$$

since these two expressions must be equal we are left with the functional equation (note however the comments in section 2.3.4)

$$F\big[x, F(y,z)\big] = F\big[F(x,y), z\big]. \quad (2.14)$$

Letting $u = F(x,y), v = F(y,z)$, $F_1(x,y) = \frac{\partial F(x,y)}{\partial x}$, $F_2(x,y) = \frac{\partial F(x,y)}{\partial y}$ and differentiating the two terms with respect to x and y we obtain

$$F_1(x,v) = F_1(u,z)F_1(x,y) \quad (2.15)$$
$$F_2(x,v)F_1(y,z) = F_1(u,z)F_2(x,y). \quad (2.16)$$

Eliminating $F_1(u,z)$ and introducing $G(x,y) = \frac{F_1(x,y)}{F_2(x,y)}$ we obtain

$$G(x,v)F_1(y,z) = G(x,y) \quad (2.17a)$$
$$G(x,v)F_2(y,z) = G(x,y)G(y,z). \quad (2.17b)$$

Here the last equation is obtained from the first by multiplying with $G(y,z)$. Differentiating eq. (2.17a) and eq. (2.17b) with z and y respectively we obtain

$$\begin{aligned}\frac{\partial}{\partial z}G(x,y) &= \frac{\partial}{\partial z}\left[G(x,v)F_1(y,z)\right] \\ &= G_2(x,v)F_2(y,z)F_1(y,z) + G(x,v)F_{12}(y,z) \\ \frac{\partial}{\partial y}G(x,y)G(y,z) &= \frac{\partial}{\partial y}\left[G(x,v)F_2(y,z)\right] \\ &= G_2(x,v)F_1(y,z)F_2(y,z) + G(x,v)F_{21}(y,z).\end{aligned}$$

Assuming $F_{12} = F_{21}$ (ie. the order of differentiation can be interchanged), the right-hand side of the two equations are equal and the first equation equal zero thus $G(x,y)G(y,z)$ is independent of y. The most general differentiable function which satisfy this condition is

$$G(x,y) = r\frac{H(x)}{H(y)} \quad (2.18)$$

for a constant r. Inserting this expression into eq. (2.17) we obtain

$$F_1(y,z) = \frac{H(v)}{H(y)} \qquad F_2(y,z) = r\frac{H(v)}{H(z)} \quad (2.19)$$

Consider small variations in y, z and $v = F(y,z)$, ie. assume y, z, v vary as $\tau \mapsto y(\tau), \tau \mapsto z(\tau), \tau \mapsto v(\tau)$. Using eq. (2.19) we obtain the differential equation

$$\frac{dv}{d\tau} = \frac{1}{dv}F(y,z) \tag{2.20}$$

$$= H(v)\left(\frac{1}{H(y)}\frac{dy}{d\tau} + r\frac{1}{H(z)}\frac{dz}{d\tau}\right) \tag{2.21}$$

introducing the function $\phi(x) = \int_{x_0}^{x} dx \frac{1}{H(x)}$ for an arbitrary (fixed) x_0 the above equation has the solutions

$$\phi(v) = \phi\big(F[y,z]\big) = \phi(y) + r\phi(z) + k_0 \tag{2.22}$$

for all constant k_0. Returning to our original expression eq. (2.14), $F[x, F(y,z)] = F[F(x,y), z]$ and applying eq. (2.22) twice on both sides we obtain

$$\phi(x) + r\phi(y) + r^2\phi(z) + 2k_0 = \phi(x) + r\phi(y) + r\phi(z) + 2k_0 \tag{2.23}$$

Assuming $\phi(z) \neq 0$ the only solution is $r = 1$. Accordingly the function F satisfies

$$\phi\big(F[x,y]\big) = \phi(x) + \phi(y) \tag{2.24}$$

Next consider three propositions A, B, C. The above equation implies the relationship between the state of beliefs $(AB|C), (B|C)$ and $(A|BC)$:

$$\phi[(AB|C)] = \phi[(B|C)] + \phi[(A|BC)]. \tag{2.25}$$

Since this relationship must hold in general it must also hold when A and C are logically equivalent. For instance we might consider the case where A corresponds where the side facing up on an ordinary dice is even and C correspond to the case where the side facing down is odd. Since these two statements are equivalent (recall the sum of opposite sides on a die is seven) it must hold that $(AB|C) = (B|C)$ and $(A|BC) = (A|C)$ and eq. (2.25) reduces to (omitting the double parenthesis, $\phi(A|B) \equiv \phi\big((A|B)\big)$)

$$\phi(B|C) = \phi(B|C) + \phi(A|C) \tag{2.26}$$

Since B could be any other proposition, for instance *"There was once life on Mars"*, and assuming ϕ takes other values than $\pm\infty$, this implies our state of belief for propositions A logically implied by other propositions C we have assumed to be true is characterized by $\phi(A|C) = 0$.

Now consider the opposite case where some proposition A' is *known* to be false given information C. For instance C corresponds to the case where the side

facing down on an ordinary dice is odd and A' is the case where the side facing up is odd too. In this case assuming C we know A' is false – accordingly $A'B$ must be false too and so $(A'B|C) = (A'|C)$. Similarly B does not change the falseness of A' given C and so $(A'|BC) = (A'|C)$. In this case eq. (2.25) reduces to

$$\phi(A'|C) = \phi(B|C) + \phi(A'|C) \tag{2.27}$$

for arbitrary B. The only way to make sense of this equation is assuming $\phi(A'|C)$ is either plus or minus infinite. While both choices are possible, we will assume $\phi(A'|C) = -\infty$. The special states of belief 0 and $-\infty$ will be denoted as *certainty*. Finally we will introduce a change of coordinates $w(x) \equiv \exp(\phi(x))$. Notice for tautologically true or false proposition w is then 1 and 0 respectively and eq. (2.25) becomes

$$w(AB|C) = w(A|BC)w(B|C). \tag{2.28}$$

2.1.3 Relationship between a proposition and its negation

Next we try to relate the belief in a statement A to our belief in it's negation $\overline{A}$ (in both situations conditional on background knowledge C, that is we are relating $A|C$ to $\overline{A}|C$). Since $\overline{A}$ determines A, it seem the least restrictive assumption is to assume there is a function S such that $S(w(A|C)) = w(\overline{A}|C)$. Notice this requirement relates to the crucial desiderata (I), that is the belief in a proposition is a *single* number. If $w(A|C)$ did not determine $w(\overline{A}|C)$ we would invariantly obtain a two-dimensional theory where beliefs was characterized by the degree of belief in $A|C$ and $\overline{A}|C$ [Shafer et al., 1976], and it is also the exclusion of such a relationship between our certainty in a proposition and it's negation which set the present theory apart from e.g. multivariate logics [Hájek, 1998].

However assume there exists such a function S. Since $\overline{\overline{A}} = A$ it must hold

$$w(A|C) = w(\overline{\overline{A}}|C) = S(w(\overline{A}|C)) = S(S(w(A|C))) \tag{2.29}$$

and thus

$$S(S(u)) = u, \qquad u = w(A|C). \tag{2.30}$$

Next consider three propositions A, B, C. Recall the distributive law for logical disjunction: $\overline{A + B} = \overline{A}\ \overline{B}$ and in particular using the distributive law: $C(\overline{CD}) = C(\overline{C} + \overline{D}) = C\overline{D}$. Now consider the two equivalent ways to compute

the belief of $(AB|C)$:

$$\begin{aligned} w(AB|C) &= w(A|C)w(B|AC) &&= w(A|C)S\left[w(\overline{B}|AC)\right] \\ &= w(A|C)S\left[\frac{w(\overline{B}|AC)w(A|C)}{w(A|C)}\right] &&= w(A|C)S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right]. \end{aligned} \tag{2.31}$$

Next, notice $AB = A(\overline{A} + B) = A\overline{A\overline{B}} = A\overline{U}$ with $U \equiv A\overline{B}$ and that $\overline{A}\ \overline{U} = \overline{A}\overline{A\overline{B}} = \overline{A}(\overline{A} + B) = \overline{A}$ regardless of B. We can then re-write:

$$\begin{aligned} \phi(AB|C) &= w(A\overline{U}|C) \\ &= w(A|\overline{U}C)w(\overline{U}|C) &&= S(w(\overline{A}|\overline{U}C))w(\overline{U}|C) \\ &= S\left[\frac{w(\overline{A}|\overline{U}C)w(\overline{U}|C)}{w(\overline{U}|C)}\right] w(\overline{U}|C) &&= S\left[\frac{w(\overline{A}\ \overline{U}|C)}{w(\overline{U}|C)}\right] \phi(\overline{U}|C) \\ &= S\left[\frac{S(w(A|C))}{S(w(A\overline{B}|C)}\right] S\left[w(A\overline{B}|C)\right]. \end{aligned} \tag{2.32}$$

Introducing $x = w(A|C)$ and $y = w(A\overline{B}|C)$ and combining eq. (2.31) and eq. (2.32) gives the functional relationship

$$xS\left[\frac{S(y)}{x}\right] = yS\left[\frac{S(x)}{y}\right]. \tag{2.33}$$

2.1.4 Solving the functional equation for negation

The function relationship in eq. (2.33) is easiest solved by assuming it is twice differentiable and non-increasing. Let $u \equiv \frac{S(y)}{x}$ and $v \equiv \frac{S(x)}{y}$. Differentiating vrt. x and y gives the following expressions

$$xS(u) = yS(v) \tag{2.34a}$$

$$\frac{\partial}{\partial x}: \quad uS'(u) - S(u) = -S'(v)S'(x) \tag{2.34b}$$

$$\frac{\partial}{\partial y}: \quad -S'(u)S'(y) = vS'(v) - S(v) \tag{2.34c}$$

$$\frac{\partial^2}{\partial x \partial y}, \frac{\partial^2}{\partial y \partial x}: \quad \frac{u}{x}S''(u)S'(y) = \frac{v}{y}S''(v)S'(x). \tag{2.34d}$$

Multiplying the right and left-hand side of eq. (2.34a) with eq. (2.34d) we obtain

$$uS''(u)S(u)S'(y) = vS''(v)S(v)S'(x). \tag{2.35}$$

Solving eq. (2.34b) and eq. (2.34c) for $S'(x), S'(y)$ respectively and substituting into eq. (2.35) we obtain

$$\frac{uS''(u)S'(u)}{(uS'(u) - S(u))S'(u)} = \frac{vS''(v)S'(v)}{(vS'(v) - S(v))S'(v)}. \tag{2.36}$$

Recall this hold for arbitrary values of A, B and C. It is easy to verify u and v can take different values in general, and so for the relationship to hold for all values of A, B and C the left-hand side and right-hand side of eq. (2.36) must be constant. Denoting this contant by k we get

$$\frac{S''(u)}{S'(u)} = k\left(\frac{S'(u)}{S(u)} - \frac{1}{u}\right) \tag{2.37}$$

which can be re-written as

$$\frac{\partial}{\partial u}\log S'(u) = k\frac{\partial}{\partial u}\left(\log S(u) - \log(u)\right). \tag{2.38}$$

Integration gives $S'(u) = a_0\frac{S(u)^k}{u^k}$ and so $\frac{\partial}{\partial u}\left(S(u)^{-k+1}\right) = a_0u^{-k}$. It is easily verified by substitution the solution of this differential equation is

$$S(x) = \left(a_0x^m + b_0\right)^{\frac{1}{m}} \tag{2.39}$$

for $m \equiv 1-k$ and constants a_0, b_0. Since $S(S(x)) = x$ we must have $a_0^2 = 1$ and $b_0 + a_0b_0 = 0$. Consider for a moment the solution $a_0 = 1$. In this case $b_0 = 0$ and so $S(x) = x$. Inserting into eq. (2.33) results in $x\frac{y}{x} = y\frac{x}{y}$, excluding this possibility. This shows only the solution $a_0 = -1$ is feasible and so we obtain the relationship

$$S(x) = \left(1 - x^m\right)^{\frac{1}{m}}. \tag{2.40}$$

Using the identity $S(S(x)) = x$ for an arbitrary beliefs $w(A|B)$ we arrive at $w(A|B) = (1 - w(\overline{A}|B)^m)^{\frac{1}{m}}$. This implies

$$w(A|B)^m + w(\overline{A}|B)^m = 1 \tag{2.41}$$

2.1.4.1 Numerical Values

In eqs. (2.26) and (2.27) it was argued somewhat informally known true and false propositions should have a belief of 0 and 1. As a consistency check this result can also be derived more easily from eq. (2.31) by letting $A = B$ and assume a general A to obtain the identity

$$w(AA|C) = w(A|C)S\left[\frac{w(A\overline{A}|C)}{w(A|C)}\right] \tag{2.42}$$

Let "False" be the false truth-value. Using eq. (2.40) for general A we arrive at

$$0 = w(\text{False}|C) \tag{2.43}$$

Thus tautologically false propositions have a degree of belief of zero and, by virtue of eq. (2.40), tautologically true statements have a degree of belief of 1. To finish the development, notice if w is a function satisfying eq. (2.28) so will w^n for any positive n. Accordingly we can define a new function $p(x) = w^m(x)$. Written using this function the product rule eq. (2.28) and the sum rule eq. (2.41) becomes

$$\begin{aligned} p(AB|C) &= p(B|C)p(A|BC) \\ &= p(A|C)p(B|AC) \end{aligned} \tag{2.44a}$$

$$p(A|C) + p(\overline{A}|C) = 1. \tag{2.44b}$$

Thus, if one accepts the desiderata and the derivation, one can conclude that

> *If a system of beliefs of binary propositions fullfill desiderata (I),(II),(IIIa),(IIIb),(IIIc) the beliefs must, up to a rescaling, follow the basic rules of probability theory* (2.44).

This result was first derived axiomatically and under comparable assumption by Cox [1946]. An extended discussion is given in Cox [1961] and as mentioned our presentation follows very closely that of Jaynes [2003]. In section 2.3 we will discuss the connection between probabilities as discussed so far and other treatments of probability theory, in particular the use of probability theory in modern Bayesian theory. However as to not get lost in a formal (and not fully resolved) discussion we will discuss three examples which highlight how the theory up to this point can be applied in different situations.

2.2 Examples

The derivation in the previous section left us with the rules of probability theory eq. (2.44) as well as numeric value of p for two beliefs, namely our belief of what is tautologically true and false: $p(\text{True}|C) = 1 - p(\text{False}|C) = 1$. This is of course a far cry from any practical applications and the following three short examples will highlight various aspects of probability theory.

2.2.1 Equivalent states of beliefs

The first example intend to answer what appears to be trivial questions: *what is our belief a flip of a coin will come up heads?*, *what is our belief a roll of a dice*

come up three? These problems might seem too trivial to even consider, however nothing derived so far can answer the questions and the answer is not entirely straight-forward. Furthermore, the assignment of states of beliefs will be the consideration of chapter 3 and so the coin and dice will serve as an important introductory exercise. Finally, the *ability* of the present theory to allow such a derivation is arguably one of the most important factors that set the theory apart from other notions of vagueness such as fuzzy theories. The derivation closely follows that of Jaynes [2003].

2.2.1.1 Mutually exclusive propositions

Consider first three propositions A, B and C and notice the following holds in ordinary propositional logic

$$A + B = \overline{(\overline{A}\ \overline{B})}. \tag{2.45}$$

This implies by repeated applications of eqs. (2.44a) and (2.44b):

$$\begin{aligned} p(A+B|C) &= 1 - p(\overline{A}\ \overline{B}|C) && = 1 - p(\overline{A}|\overline{B}C)p(\overline{B}|C) \\ &= 1 - \left[1 - p(A|\overline{B}C)\right] p(\overline{B}|C) && = p(B|C) + p(A\overline{B}|C) \\ &= p(B|C) + p(\overline{B}|AC)p(A|C) && = p(B|C) + \left[1 - p(B|AC)\right] p(A|C) \\ &= p(A|C) + p(B|C) - p(AB|C) \end{aligned} \tag{2.46}$$

For the general case, consider n propositions $A_1, \ldots, A_n$. As suggested by eq. (2.46) we now prove by induction over n that

$$p(A_1 + \cdots + A_n|C) = \sum_{k=1}^{n} (-1)^{k+1} \left[\sum_{1 \le i_1 < \cdots < i_k \le n} p(A_{i_1} \ldots A_{i_k}|C) \right] \tag{2.47}$$

in particular letting $A = A_n$ and $B = A_1 + \cdots + A_{n-1}$ and using eq. (2.46) we have $p(A + B|C) = p(A|C) + p(B|C) - p(B|AC)p(A|C)$. By the induction

assumption eq. (2.47)

$$\begin{aligned}
p(A_1+\cdots+A_n|C) &= p(A_n|C)+\sum_{k=1}^{n-1}(-1)^{k+1}\left[\sum_{1\leq i_1<\cdots<i_k\leq n-1} p(A_{i_1}\ldots A_{i_k}|C)\right] \\
&\quad - p(A_n|C)\sum_{k=1}^{n-1}(-1)^{k+1}\left[\sum_{1\leq i_1<\cdots<i_k\leq n-1} p(A_{i_1}\ldots A_{i_k}|A_nC)\right] \\
&= \sum_{i=1}^{n} p(A_n|C)+\sum_{k=2}^{n-1}(-1)^{k+1}\left[\sum_{1\leq i_1<\cdots<i_k\leq n-1} p(A_{i_1}\ldots A_{i_k}|C)\right] \\
&\quad + \sum_{k=2}^{n}(-1)^{k+1}\left[\sum_{1\leq i_1<\cdots<i_{k-1}<i_k=n} p(A_{i_1}\ldots A_{i_k}|C)\right] \\
&= \sum_{k=1}^{n}(-1)^{k+1}\left[\sum_{1\leq i_1<\cdots<i_k\leq n} p(A_{i_1}\ldots A_{i_k}|C)\right] \qquad (2.48)
\end{aligned}$$

which proves the result.

Suppose the information contained in C implies no two of the hypothesis A_i and A_j can be true at the same time, for instance the hypothesis may be which side of a n-sided dice faces upwards. In this case

$$p(A_iA_j|C)=\delta_{ij}p(A_i|C). \qquad (2.49)$$

Plugging this into eq. (2.47) all terms with more than two elements vanish and we arrive at the general result that for all subsets $1\leq i_1<\cdots<i_k\leq n$:

$$p(A_{i_1}+\cdots+A_{i_k}|C)=p(A_{i_1}|C)+\cdots+p(A_{i_k}|C). \qquad (2.50)$$

Next, assume in addition to eq. (2.49) that the background information C ensures one of $A_1,\ldots,A_n$ is true. In this case $A_1+\cdots A_n$ is a tautology and must have probability 1 from the discussion in the previous chapter. As a result we obtain

$$\sum_{i=1}^{n} p(A_i|C)=1. \qquad (2.51)$$

2.2.1.2 Numerical values

So far we have still not arrived at any definite numerical values for beliefs, only shown if the propositions $A_1,\ldots,A_n$ are rendered exhaustive and exclusive on

the background information C eq. (2.51) must hold. To arrive at numerical values we will use the two desiderata (IIIb) and (IIIc). The analysis is mathematical trivial, but the argument requires some care. To avoid confusion we will therefore use an example of a dice. Suppose we are considering a roll of a dice with $n = 6$ faces and the 6 propositions $A_1, \ldots, A_n$:

$$\begin{aligned} A_1 &: \textit{The side } \boxdot \textit{ face up} \\ A_2 &: \textit{The side } \boxdot \textit{ face up} \\ A_3 &: \textit{The side } \boxdot \textit{ face up} \\ A_4 &: \textit{The side } \boxdot \textit{ face up} \\ A_5 &: \textit{The side } \boxdot \textit{ face up} \\ A_6 &: \textit{The side } \boxdot \textit{ face up.} \end{aligned} \tag{2.52}$$

In this case we assume the propositions are exhaustive and independent such that eq. (2.51) holds. Suppose we now consider two problems. In problem (I) the propositions are labelled as in eq. (2.52) and give rise to an assignment of belief $p_I(A_i|C)$ to each proposition. The second problem (II) considers a similar set of propositions $A'_1, \ldots, A'_n$ defined as a permuted version of the first, for instance by interchanging $\boxdot$ and $\boxdot$:

$$\begin{aligned} A'_1 &: \textit{The side } \boxdot \textit{ face up} \\ A'_2 &: \textit{The side } \boxdot \textit{ face up} \\ A'_3 &: \textit{The side } \boxdot \textit{ face up} \\ A'_4 &: \textit{The side } \boxdot \textit{ face up} \\ A'_5 &: \textit{The side } \boxdot \textit{ face up} \\ A'_6 &: \textit{The side } \boxdot \textit{ face up} \end{aligned} \tag{2.53}$$

and this again give rise to an assignment of beliefs $p_{II}(A_i|C)$. Since the first example only differs from the second in the labelling of the propositions, the assignment of belief must be consistent for the two first propositions. In particular

$$\begin{aligned} p_I(A_1|C) &= p_{II}(A'_2|C) \\ p_I(A_2|C) &= p_{II}(A'_1|C) \\ p_I(A_i|C) &= p_{II}(A'_i|C) \text{ for } i \geq 3. \end{aligned} \tag{2.54}$$

Clearly this relationship holds whatever C might be and whatever the numerical values may be. We now arrive at a subtle but crucial point: We suppose the information C is *indifferent* with regards to the propositions A_1 and A_2 in eq. (2.52). Notice this is *not* to say the dice is *unweighted*, *fair*, *of equal dimension and uniform density* or that it was throw *randomly* or *from an arbitrary initial position* or *from a sufficient height and velocity* or *by a fair player* or some other incantation often used in these situations: It is difficult to say what these terms mean and it is natural to suppose no dices are of equal dimension

or density. Rather, the statement mean the totality of information in C does not contain any statements which force us to believe more strongly proposition A_1 over A_2 or visa-versa. Where this may be somewhat counter-intuitive is if we (for instance) know the number in a dice are often made by indentation and this means the density near the side ⚅ is often lower than near ⚀; we need to suppose such information is not contained in C.

Suppose C is indifferent with respect to A_1 and A_2. In this case we should suppose the assignment of beliefs in problem I is equivalent to that in problem II regardless of the labelling and so by desiderata (IIIc)

$$p_I(A_i|C) = p_{II}(A'_i|C) \text{ for all } i. \tag{2.55}$$

Comparing eq. (2.54) and eq. (2.55) we arrive at

$$p_I(A_1|C) = p_I(A_2|C) \tag{2.56}$$

ie. the assignment of belief to ⚀ and ⚁ is the same. Applying this argument for other propositions we arrive at our result:

$$p_I(A_i|C) = p_{II}(A'_i|C) = \frac{1}{n} \tag{2.57}$$

no doubt more useful than surprising.

2.2.2 The Tuesday paradox

The second example relates to the so-called Tuesday paradox [Gardner, 1959]. Simply stated:

> *A man has two children. One is a girl born on a Tuesday. How many daughters do the man have?* (2.58)

To solve this riddle, we need to translate it into appropriate propositions. The first relevant proposition is what we are interested in, namely if the man have to daughters. In addition to this we also need to encode the information given in the text. Let the two children be denoted by 1 and 2 (for instance the order in which they were born). The following set of six propositions are sufficient:

C : *The man have two children*

B : *It is not the case that 'the man have a child which is a girl born on a tuesday'*

G_1 : *Child* 1 *is a girl*

G_2 : *Child* 2 *is a girl*

T_1 : *Child* 1 *was born on a Tuesday*

T_2 : *Child* 2 *was born on a Tuesday*

In this language, the information we are given is C and $\overline{B}$, and we are interested in computing

$$p(G_1G_2|\overline{B}C) \tag{2.59}$$

Using the product and sum rules eq. (2.59) becomes

$$\begin{aligned} p(G_1G_2|\overline{B}C) &= \frac{p(\overline{B}|G_1G_2C)p(G_1G_2|C)}{p(\overline{B}|C)} \\ &= \frac{\sum_{t_1,t_2} p(\overline{B}|G_1G_2t_1t_2C)p(t_1t_2|G_1G_2C)p(G_1G_2|C)}{\sum_{t_1t_2g_1g_2} p(\overline{B}|t_1t_2g_1g_2C)p(t_1t_2g_1g_2|C)}. \end{aligned} \tag{2.60}$$

The notation $\sum_{t_1} p(At_1|C)$ is shorthand for $p(AT_1|C) + p(A\overline{T_1}|C)$. When we know the sex and day of birth of both children $\overline{B}$ will be known true or false. The remaining terms become

$$\begin{aligned} p(G_1G_2|\overline{B}C) = \frac{p(T_1\overline{T_2}G_1G_2|C) + p(\overline{T_1}T_2G_1G_2|C) + p(T_1T_2G_1G_2|C)}{2(p(T_1\overline{T_2}G_1\overline{G_2}|C) + p(T_1T_2G_1\overline{G_2}|C))} \\ +(p(T_1\overline{T_2}G_1G_2|C) + p(\overline{T_1}T_2G_1G_2|C) + p(T_1T_2G_1G_2|C)) \end{aligned} \tag{2.61}$$

To proceed requires numerical values. We obtain these by assuming (1) knowing the sex of a child tell us nothing about the day of week the child was born (2) knowing the sex or day of the week one child is born tell us nothing nothing about the sex or day of the week of the other, ie.

$$p(t_1t_2g_1g_2|C) = p(t_1|C)p(t_2|C)p(g_1|C)p(g_2|C). \tag{2.62}$$

Furthermore we assume there is no preferred sex or day of the week, ie. our reasoning would be the same if the problem was stated with Monday instead of Tuesday, boy instead of girl. In this case the analysis of equal states of belief of section 2.2.1 holds and we obtain $p(G_1|C) = \frac{1}{2}$ and $p(T_1|C) = \frac{1}{7}$. Then eq. (2.61) becomes

$$\begin{aligned} p(G_1G_2|\overline{B}C) &= \frac{\frac{6}{7^2} + \frac{6}{7^2} + \frac{1}{7^2}}{2\left(\frac{6}{7^2} + \frac{1}{7^2}\right) + \left(\frac{6}{7^2} + \frac{6}{7^2} + \frac{1}{7^2}\right)} \\ &= \frac{13}{27}. \end{aligned} \tag{2.63}$$

It should however be noticed this is only under the assumption the information in (2.58) was correctly captured by the above translation into Boolean statements and their relationship, see Falk [2011] for a longer discussion of this point. The name Tuesday paradox stems from possibly counter-intuitive nature of eq. (2.63). Intuitively one might assume by the independence assumption

eq. (2.62) that the correct probability is 1/3 (there are four combinations of boy/girl for the two children and we can rule out the option of two boys), however that it is the child *which is known to be a girl* that is born on a Tuesday means these options are not symmetric. Finding out what this effect is in a quantitative manner is difficult using a counting argument, however by using Bayes theorem it is fairly easily translated into simpler questions.

2.2.3 Lotteries and Jesus

The final example illustrates how Bayes theorem can be used in a qualitative manner to assess the goodness of an argument. It is inspired by a passage in Strobel [2004] where the philosopher Dr. William Lane Craig discusses the reasonableness of being sceptical of what some might otherwise consider an unlikely proposition. In the book, Lee Strobel asks Dr. Craig the following question regarding the resurrection of Jesus:

> *"Some critics say that the Resurrection is an extraordinary event and therefore it requires extraordinary evidence,", I [Lee Strobel] said. "Doesn't that assertion have a certain amount of appeal?"*
> *–"Yes, that sounds like common sense," he replied. "But it's demonstrably false."*
> *"How so?"*
> *–"This standard would prevent you from believing in all sorts of events that we do rationally embrace. For example, you would not believe the report on the evening news that the numbers chosen in last night's lottery were 4, 2, 9, 7, 8 and 3, because that would be an event of extraordinary improbability. The odds against that are millions and millions to one, and therefore you should not believe it when the news reports it."* [Strobel, 2004, p. 65].

Many people would likely be sceptical of this argument. After all, if the argument is true, the husband of the married couple introduced earlier could have asked his wife to consider the above argument too[4]: Just because the sequence of events he described the other night (the secretary who slipped, the accidental lipstick-smear, the locked-out cell phone, his sudden illness) seems *extraordinary* on the surface she should not jump to any conclusions and ask for extraordinary evidence it did happend that way.

Let us try to investigate the argument using probability theory. First we need to translate the argument into language we can recognize. The primary events

[4]Unless ones spouse is very philosophically inclined this is not advisable.

Dr. Craig asks us to consider are the events D_{s_0}, $s_0 = (4, 2, 9, 7, 8, 3)$ and J where

$$D_s : \textit{The sequence of lottery numbers } \textbf{\textit{actually}} \textit{ drawn was } s$$
$$J : \textit{Jesus } \textbf{\textit{actually}} \textit{ rose bodily from the dead in the first century.}$$

We emphasize these events correspond to what did in fact happen, not what people claim happened. Now our interest in these claims relates to our belief in their truth today given what we know. In the case of J what we have available to assess the claims truth are various historical sources such as manuscripts of the Bible. Similarly, for D_{s_0} the information we have access to is *what the newsman said*. Accordingly we introduce two additional propositions

$$N_{s_0} : \textit{The newsman reported the lottery sequence } s_0$$
$$B : \textit{The Bible report Jesus rose from the dead.}$$

It will be assumed what Dr. Craig refer to as extraordinary is extraordinary with respect to the evidence we have available. Accordingly, what we should evaluate (or compare) the extraordinariness of is then:

$$p(D_{s_0}|N_{s_0}\Omega) = \frac{p(N_{s_0}|D_{s_0}\Omega)p(D_{s_0}|\Omega)}{p(N_{s_0}|\Omega)}, \tag{2.64}$$

$$p(J|B\Omega) = \frac{p(B|J\Omega)p(J|\Omega)}{p(B|\Omega)} \tag{2.65}$$

where the proposition Ω is shorthand for all other relevant information. To proceed we need to make additional assumptions. Firstly, we will assume the chance of a person rising from the dead is approximately the same as winning the lottery, and we will assume there are $n = 10^6$ lottery sequences; I do not know what a doctor would make of this assumption, but it highlights the more important aspects of the argument.

Under these assumptions eq. (2.65) becomes

$$p(D_{s_0}|N_{s_0}\Omega) = \frac{p(N_{s_0}|D_{s_0}\Omega)p(D_{s_0}|\Omega)}{\sum_s p(N_{s_0}|D_s\Omega)p(D_s|\Omega)} \tag{2.66}$$

$$= \frac{p(N_{s_0}|D_{s_0}\Omega)}{p(N_{s_0}|D_{s_0}\Omega) + \sum_{s\neq s_0} p(N_{s_0}|D_s\Omega)} \tag{2.67}$$

where the sum is over all possible lottery sequences. We then need to consider what could possibly be the reason we might hear the wrong lottery sequence; if we assume a sequence is always drawn, we should consider the possibilities that the newsman made an error in reading the sequence, that someone transcribed it falsely, that someone swindled with the machine such that it report the wrong sequence and so on. We thus introduce the variable:

$$H : \textit{The newsreport of the sequence was honest}$$

Ie. $p(N_{s_1}|D_{s_2}H\Omega) = 1$ only if $s_1 = s_2$. However suppose there is an error in the lottery, $\overline{H}$. In this case the sequence is made up, and as a simplifying assumption we assume s_2 offers no information regarding which particular sequence the newsman should read out. Assuming sequences are drawn randomly in the lottery this amounts to $p(N_{s_1}|D_{s_2}\overline{H}\Omega) = p(D_{s_1}|\Omega)$ and $p(D_s|\Omega) = \frac{1}{n}$. Under these assumptions eq. (2.67) become

$$\begin{aligned} p(D_{s_0}|N_{s_0}\Omega) &= \frac{p(H|\Omega) + p(D_{s_0}|\Omega)p(\overline{H}|\Omega)}{p(H|\Omega) + p(D_{s_0}|\Omega)p(\overline{H}|\Omega) + \sum_{s\neq s_0} p(D_s|\Omega)p(\overline{H}|\Omega)} \\ &= p(H|\Omega) + \frac{1 - p(H|\Omega)}{n}. \end{aligned} \tag{2.68}$$

It is interesting to consider some limit cases in this expression. If the announcer is maximally untrustworthy the above reduce to $\frac{1}{n}$. On the other hand if he is maximally trustworthy it become 1. Compare this to the case of Jesus, making the assumption a resurrection is as extraordinary as a lottery sequence, $p(J|\Omega) = p(D_{s_0}|\Omega) = \frac{1}{n}$ and, assuming Jesus rose from the dead, Jesus also had the foresight to employ reliable Bible writers $p(B|J\Omega) = 1$ then eq. (2.65) reduces to

$$\begin{aligned} p(J|B\Omega) &= \frac{p(B|J\Omega)p(J|\Omega)}{p(B|J\Omega)p(J|\Omega) + p(B|\overline{J}\Omega)p(\overline{J}|\Omega)} \\ &= \frac{1}{1 + p(B|\overline{J}\Omega)(n-1)}. \end{aligned} \tag{2.69}$$

The qualitative difference between eq. (2.68) and eq. (2.69) should be apparent. If we assume n is large, *all* that matters when determining if the reported sequence really did get drawn is how honest we believe the newsreporter is, not the number of possible sequences. On the other hand for the resurrection story to be believable it need to be argued the chance of there arising a *myth* someone rose from the dead should be about equally extraordinary as someone actually rising from the dead, given the same background information. We can then conclude the two examples are not equivalent since, on the assumption the lottery sequence is reported dishonestly, there is the same chance of the particular sequence being reported as there would be if the sequence was reported honestly.

Most people, regardless of their beliefs in the resurrection of Jesus, intuitively guess the situation of the lottery and the risen Christ are not entirely equivalent. The advantage of probability theory is that the dissimilarities can be made more apparent and the type of assumption one needs to make to argue they remain equivalent becomes more apparent.

2.3 From basic probability theory to probability theory

The foundations of probability theory has historically been a controversial subject centered around the interpretation of probabilities. Why should we care about that today? A practical answer is the derivation so far which left us with the basic rules of probability theory eq. (2.44) defined on binary propositions suffers from the flaw of being an inadequate foundation for most of Bayesian non-parametrics. This difficulty will be the subject of the following sections.

2.3.1 A brief history of probability

From the dawn of ages hunter-gatherers, farmers and gambling city-dwellers have been concerned with judgements regarding the behaviour of prey, the weather and the outcome of bets and games. Randomness and uncertainty has therefore always been a part of every-day experience. It is on this view surprising that probability was first scrutinized relatively late and for the most of human history probabilities and randomness was treated informally. For instance in the writings of the ancient Greeks such as the Atomists Epicurus and Lucretius (ca. 300 and 50BC) believed the world to be composed of swerving "atoms" whose behaviour was random but governed by laws [Russell, 1946] and Aristotle provided the following account in Rhetoric, ca. 350BC:

> *the probable is that which for the most part happens.*
> –English translation by Roberts et al. [1954]

However none of this work work involved an explicit study of probability extending much beyond common-sense intuitions.

The earliest modern accounts of probability can be traced back to around 1654 in a letter exchange between Pierre de Fermat and Étienne Pascal [De Fermat and Pascal, 1654]. In the exchange they considered the outcome of games such as when it is a loosing strategy to bet on certain outcomes in repeated throws of a dice. The next significant development is associated with Bayes and Price [1763] and de Laplace [1820]. Bayes is credited for introducing the concept of conditional probabilities and realizing these can be expressed in terms of other probabilities $p(X|Y) = p(Y|X)p(X)/p(Y)$, however his work was significantly expanded by the mathematician Laplace who is responsible for the mathematical development of these and other ideas. Laplace offered the following definition of probabilities:

> *The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.*
> –Laplace 1814, english translation due to [Laplace and Dale, 1995]

This definition no doubt captures the common-sense intuition behind probabilities and fares well when one for instance considers draws from a game of cards or rolls of a dice. However if we for instance considering arguably one-off events (such as if it will rain tomorrow) the definition seems to lead to counter-factual statements. In addition it has been argued it is difficult to formalize the notion of *"equally possible"* without introducing probabilities in some way; as an additional technical annoyance the definition would seemingly be unable to account for non-rational probabilities such as $1/\sqrt{2}$. See Jaynes [2003] for further discussion. If we focusing on the difficulties which arise from the phrase *"equally possible"* in the above definition there are two main avenues one can explore which persists until today: [Jaynes, 2003, Corfield and Williamson, 2001]

The frequentist view: Which hinge the definition of probabilities on the outcome of repeated events which we will denote the *frequentist view*. Crudely stated, is is that to say e.g. a coin has probability 0.52 of coming up heads is to say the limit frequency of times the coin come up heads to the total number of flips converges towards 0.52. This number then reflects a *physical* property of the irregularities of the coin and is said to be *objective* in that another person could arrive at the same number though another sequence of coin flips.

The subjectivist view: In which probabilities are associated with a *state-of-knowledge* of a rational agent about events which are not thought to necessarily re-occur. Probabilities are either assigned through symmetry considerations or derived from other probabilities according to the rules of probability theory. The preceding sections has been an example of a definition of probabilities which follows the subjective view and this will also be designated the *Bayesian view* of probabilities.

The frequentist and subjectivist view crystalized in the 19th century and towards the beginning of the 20th century. In the first half of the 19th century a number of mathematicians amongst these Poisson [1837], Cournot [1843] and Ellis [1843]

introduced many important probabilistic concepts and methods and discussed the distinction between the frequentist and subjectivist view with an increasing preference towards the former. Near the end of the 19th century the frequentist view became predominant and for instance Venn [1866] identifies probabilities with a frequentist interpretation. For our purpose however the most important contribution to the frequentist view was the mathematically rigorous foundation of the frequentist view of probabilities provided by Kolmogorov [1933] which will be discussed further below.

Turning to the subjectivist view of probabilities, while subjectivist ideas can be traced back to the beginning of the modern study of probabilities, the modern crystallization of the subjective/Bayesian view took place in the first half of the 20th century. An important treatment was given by de Finetti [De Finetti, 1937, de Finetti, 1974]. His approach, which will be discussed further below, was based on expectations of a rational agent engaged in a betting game. As mentioned many times, the approach view discussed in this thesis follows Cox [1946] [Cox, 1961] and Jaynes [2003] and is based on analysing natural-language *desiderate* a and then derive consistency requirements a rational assignment of degrees-of-beliefs must obey. It is in this respect worth mentioning the important historical treatment of probabilities as a general tool for inductive reasoning on (amongst other things) mathematical theorems by Polya [1954].

This leaves us with three foundations for probability theory: The frequentistic view associated with Kolmogorov [1933], and the two subjectivist/Bayesian views motivated with the assignment of rational betting odds approach of De Finetti [1937] and the treatment of rational degrees-of-belief of Cox [1946] and Jaynes [2003]. In the following these views will be elaborated so as to explore what *practical, mathematical tools they provide* and not from the perspective of their philosophical underpinnings. A reader interested in a more comprehensive treatment of the interpretations of probability theory and their relationship is referred to Jaynes [2003] (and references therein) as well the historical treatments of [Corfield and Williamson, 2001, Van Horn, 2003] and Fienberg et al. [2006].

2.3.2 The Kolmogorov account of probabilities

Andrey Kolmogorov was a mathematician and his approach to probability theory is in the main concerned with establishing a rigorous foundation on which frequentist probability theory can be build rather than providing a philosophical interpretation of probabilities. However to motive his idea, suppose we consider a general set $\mathcal{X}$ (keep for instance the example where $\mathcal{X} \subseteq \mathbb{R}^d$ in mind). Suppose some procedure selects elements $x_1, x_2, \ldots$ from $\mathcal{X}$ at random, independently

but not necessarily uniformly (the phrases are understood loosely). For instance suppose $\mathcal{X} = \mathbb{R}$ and the events are the length of fish captured in a lake, common sense dictates the length of the fish will *tend to* fall within a certain interval and never be negative.

If we then consider a fixed subset of $E \subseteq \mathcal{X}$ we can then consider the sequence of binary events that $x_1, x_2, \ldots$ fall within E:

$$x_1 \in E,\ x_2 \in E,\ x_3 \in E, \ldots \tag{2.70}$$

then, if we wish to use some function p_K to define the *probability* of this event, i.e. loosely stated

$$p_K(E) = \{\text{Probability the next element fall within } E\} \tag{2.71}$$

then this function should fulfill certain natural conditions. For instance $p_K(\mathcal{X}) = 1$, p_K should be non-negative and suppose $E_1, E_2 \subset \mathcal{X}$ and $E_1 \cap E_2 = \emptyset$ then

$$p_K(E_1 \cup E_2) = p_K(E_1) + p_K(E_2). \tag{2.72}$$

For countable finite sets $\mathcal{X}$ this construction can be made rigorous without any problems, however Kolmogorov observed that when $\mathcal{X} \subseteq \mathbb{R}^d$ one quickly runs into problems. Consider for instance the Banach-Tarski pardox [Banach and Tarski, 1924, Tao, 2011] which states that, assuming the axiom of choice is true, then given the unit ball $\mathcal{X} = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 \leq 1\}$ there exists a partition of $\mathcal{X}$:

$$\mathcal{X} = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5, \text{ and } A_i \cap A_j = \emptyset \text{ for } i \neq j \tag{2.73}$$

such that if suitable translations and rotations are applied to the sets A_1, A_2 and A_3, A_4, A_5 they can be put together again to form two unit balls identical to $\mathcal{X}$, in clear violation of how any notion of *uniform* probability on $\mathcal{X}$ should behave. Kolmogorovs solution was to limit p_K to only be defined on certain subsets of $\mathcal{X}$ denoted a σ-algebra.

Definition 2.3.1 (σ-algebra)**.** *A σ-algebra $\mathfrak{F}$ of a set $\mathcal{X}$ is a subset of the powerset of $\mathcal{X}$ such that: (i) $\mathcal{X} \in \mathfrak{F}$ (ii) If $A \in \mathfrak{F}$ then $\mathcal{X} \setminus A \in \mathfrak{F}$ (iii) If $A_1, A_2, \ldots$ is a countable collection of elements of $\mathfrak{F}$ then $A_1 \cup A_2 \cup \cdots \in \mathfrak{F}$.*

The smallest σ-algebra containing the open sets of $\mathcal{X}$ is also known as the Borel-algebra. With this in place we can introduce the Kolmogorov concept of probabilities

Definition 2.3.2 (Kolmogorov probability)**.** *Given a set $\mathcal{X}$ and a σ-algebra $\mathfrak{F}$ of $\mathcal{X}$ a (Kolmogorov) probability is then a function $p_K : \mathfrak{F} \to [0, 1]$ fulfilling for*

all $E \in \mathfrak{F}$ *and all disjoint countable collections* $E_1, E_2, \cdots \in \mathfrak{F}$

$$\textit{Normalization:} \qquad p_K(\mathcal{X}) = 1 \tag{2.74a}$$

$$\textit{Non-negativity:} \qquad p_K(E) \geq 0 \tag{2.74b}$$

$$\textit{Countable additivity:} \qquad p_K\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} p_K(E_i). \tag{2.74c}$$

Since Kolmogorovs definition of probability requires the σ-algebra to be specified it is common to specify $(\mathcal{X}, \mathfrak{F}, p_K)$ also denoted the *probabilistic triplet*.

It is worth noting that Kolmogorovs definition of probabilities departs in three ways from the probability p of eq. (2.44). Firstly, by virtue of being defined on sets and not a Boolean algebra of propositions. Secondly, by being defined as a function of a single *event* and not a conditional distribution $p(A|B)$ and thirdly, by countable additive eq. (2.74c).

2.3.3 The de Finetti account of probabilities

De Finetti rejected the idea probabilities should or could reflect an objective property of the world but subscribed to a subjectivist view where probabilities, as for C.T. Cox, reflected degree of belief. The result obtained by de Finetti will be very similar to that of Cox discussed in the introduction and we will therefore avoid the formal treatment to the next section. De Finetti defined probability as relating to our propensity for placing bets on propositions in light of certain information. In particular de Finetti consider the sets of available propositions to be of the same form as those discussed in the introduction and considered

then considered the probability to be a function $p_F(A|B) \mapsto q \in [0, 1]$ specifying our propensity of placing a bet on proposition A given information B (compare to the derivation in the preceding section). This propensity of placing bets was then analysed under the requirement of being suitable for monetary betting in an idealized situation known as a *Dutch book* argument [De Finetti, 1937, de Finetti, 1974].

A Dutch book argument itself begins with the *Dutch Book theorem*, which describe the conditions under which a set of bets of a particular form guarantees a net loss to one side, i.e. a Dutch Book [5]. Following de Finetti, for a proposition A it is a bet which takes the form indicated in table 2.1. The table indicates the payoff to a player who buys a bet with a stake S for the price qS such that the

[5] A Dutch book is a set of odds which guarantee profit for one side regardless of outcome

H	Payoff
True	$S - qS$
False	$-qS$

Table 2.1: Payout matrix for a bet with a stake S over a proposition A. A player buy a stake in the bet for an amount qS and receives a payout of S if A is true and otherwise nothing.

player receives an amount S if A is true and otherwise nothing. q is known as the *betting quotients* and should be computed by p_F. The Dutch book theorem then says, loosely stated, that if a set of betting quotients, i.e. p_F, fails to satisfy the probability axioms eq. (2.44) there is a set of bets with those quotients that guarantees a loss to one side.

2.3.4 The Cox account of probabilities

Finally there is the derivation of Cox [1946] discussed earlier which provided us the same end-point as the Dutch book argument of de Finetti. The probability function eq. (2.44) will in the following be denoted p_C. As mentioned, the propositions p_C is defined on form a Boolean algebra [Jaynes, 2003]. Recall the definition of a Boolean algebra is a set $\mathcal{A}$ of propositions and structure: [Birkhoff and Lane, 1977]

Definition 2.3.3 (Boolean algebra)**.** *A Boolean algebra for a set $\mathcal{A}$ is a structure $(\mathcal{A}, +, \cdot, \mathtt{f}, \mathtt{t})$ with two binary operations $+$ and $\cdot$ ("or" and "and"), a unary operation $-$ (negation), and two distinguished elements $\mathtt{f}$ and $\mathtt{t}$ ("true" and "false") such that for all $A, B, C \in \mathcal{A}$ the following holds*

$$A + (B + C) = (A + B) + C, \qquad A \cdot (B \cdot C) = (A \cdot B) \cdot C, \tag{2.75a}$$
$$A + B = B + A, \qquad A \cdot B = B \cdot A, \tag{2.75b}$$
$$A + (A \cdot B) = B, \qquad A \cdot (A + B) = A, \tag{2.75c}$$
$$A \cdot (B + C) = (A \cdot B) + (A \cdot C), \qquad A + (B \cdot C) = (A + B) \cdot (A + C), \tag{2.75d}$$
$$A + (-A) = \mathtt{t} \qquad A \cdot (-A) = \mathtt{f}. \tag{2.75e}$$

In anticipation of eq. (2.74c) we will assume the Boolean algebra in question is always closed under under countable disjunction and conjunction, i.e. if $A_1, A_2, \cdots \in \mathcal{A}$ then

$$A_1 + A_2 + A_3 + \ldots \in \mathcal{A} \tag{2.76a}$$
$$A_1 \cdot A_2 \cdot A_3 \cdot \ldots \in \mathcal{A}. \tag{2.76b}$$

and adopt the familiar notation of abbreviating $-A = \overline{A}$ and $A \cdot B = AB$ to be consistent with the notation in the introduction. Recall the semantics of the conditional belief $p_C(A|B)$ is *A given B is true*. Thus, $p_C(A|\mathtt{f})$ would be undefined. A Cox-Jaynes degree-of-belief based probability is then a function

$$p_C : \mathcal{A} \times (\mathcal{A} \setminus \{\mathtt{f}\}) \to [0, 1] \tag{2.77}$$

such that p_C satisfies that for all $A, B, C, A_i, \ldots, A_n$ such that $A_i A_j = \delta_{ij}$ and $C \neq \mathtt{f}$:

$$\text{Normalization:} \qquad p_C(\mathtt{t}|C) = 1 \tag{2.78a}$$

$$\text{Non-negativity:} \qquad p_C(A|C) \geq 0 \tag{2.78b}$$

$$\text{Finite additivity:} \qquad p_C\left(\bigcup_{i=1}^{n} A_i|C\right) = \sum_{i=1}^{n} p_C(A_i|C) \tag{2.78c}$$

$$\text{Product rule:} \qquad p_C(AB|C) = p_C(A|BC)p_C(B|C). \tag{2.78d}$$

which is just a re-write of the rules eq. (2.44) and the derivation of section 2.2.1 was used to obtain eq. (2.78c).

Comments on the derivation of p_C: A technical issue is the function p_C is not necessarily uniquely identified by the Cox desiderata. This was first noticed by Paris [1994] and extensively discussed by Halpern [1999]. The issue affects both the derivations of Cox [1946], Jaynes [2003] as well as other treatments not discussed here [Horvitz et al., 1986, Heckerman, 1986, Aleliunas, 1990] and arises when transiting from a general functional equation of the form eq. (2.13a):

$$\begin{aligned}(ABC|D) &= F\big[(BC|D), (A|BCD)\big] = F\big\{F\big[(C|D), (B|CD)\big], (A|BCD)\big\} \\ (ABC|D) &= F\big[(C|D), (AB|CD)\big] = F\big\{(C|D), F\big[(B|CD), (A|BCD)\big]\big\}\end{aligned}$$

to the statement it then holds *in general* for all x, y, z in the codomain of the assignment of degree of belief eq. (2.14):

$$F\big[x, F(y, z)\big] = F\big[F(x, y), z\big].$$

since in the former equation (which *do* hold in general) the three propositions corresponding to x, y, z was assumed to take a particular structural form:

$$x = (C|D), \quad y = (B|CD), \quad z = (A|BCD) \tag{2.79}$$

and so the second equation has only been shown to hold for those values of x, y, z *which can be formed by quintets of propositions* A, B, C, D *of that same structural form.* Halpern [1999] also discusses how this difficulty affects similar arguments such as Aczél [1966, section 7, theorem 1] and Reichenbach [1950].

Evidently this is only a potential issue when the allowed triplets (x, y, z) are restricted in a non-trivial manner by the requirement of being obtainable as the beliefs of (structured) propositions eq. (2.79). In particular the problem can be expected to crop up when the image of the degree-of-belief function $(\cdot)$ is finite and in particular when $\mathcal{A}$ is finite. Halpern [1999] constructs an explicit example for finite $\mathcal{A}$ where an assignment of degrees-of-belief satisfy the Cox axioms but does not correspond to a probability assignment in the usual sense of eq. (2.78). Paris [1994] avoids the potential problem by postulating an additional axiom for the assignment of degrees-of-belief $(\cdot)$ namely (in our notation):

Definition 2.3.4 (Paris' requirement)**.** *For all* $0 \leq a, b, c \leq 1$ *and* $\epsilon > 0$*, there are propositions* $A_1, A_2, A_3, A_4 \in \mathcal{A}$ *such that* $A_4 \Rightarrow A_3$*,* $A_3 \Rightarrow A_2$*,* $A_2 \Rightarrow A_1$*,* $A_3 \neq \boldsymbol{f}$ *and*

$$\max\{|(A_4|A_3) - a|,\ |(A_3|A_2) - b|,\ |(A_2|A_1) - c|\} < \epsilon \tag{2.80}$$

and in addition that the degree-of-belief should be contained in the interval $[0, 1]$. Notice this formulation differs from Paris [1994] in that he assumes the propositions takes the structure of a σ-algebra instead of a Boolean algebra which we will return to in a moment. Neither Paris [1994] or Halpern [1999] considers this extra requirement to be very aesthetically pleasing and at any rate the extra requirement of definition 2.3.4 requires $\mathcal{A}$ to be infinite. Naturally, for most practical implications the requirement that $\mathcal{A}$ should be infinite is not of much concern, and arguably if the propositions in $\mathcal{A}$ are sufficiently flexible to allow all (x, y, z) (or a dense subset) to be expressed the requirement definition 2.3.4 would not be required.

Asides the de Finetti derivation, the rules of probability theory eq. (2.44), eq. (2.78) has been derived from other starting points which can be roughly classified as taking the same subjective, degree-of-belief approach to probabilities as Cox [1946]. These include the more throughout treatment of Paris [1994] which is very similar to the approach of Cox and Jaynes but more throughout and relates the result to other notions of uncertainty and vagueness as well as Van Horn [2003] which also provides a variant of the derivation of Cox. Other approaches worth mentioning is the approach of Dupre and Tipler [2006] based on retraction mappings which need not assume differentiability when deriving the sum/product rule however must assume stronger conditions on the set of propositions $\mathcal{A}$. Hardy [2002] derives the sum and product rule as a special case of the more general framework of *scaled Boolean algebras*. Knuth and Skilling [2012] provides a simple proof in the case $\mathcal{A}$ is finite. Zimmermann and Cremers [2011] discusses a general, foundational approach to Cox-type derivations which includes representing beliefs as matrices or complex numbers. The subtle distinctions of these approaches cannot be surveyed here in details, however importantly none of these approaches derive the rules of probability theory with

countable additivity eq. (2.74c).

2.3.5 Comparing the Kolmogorov and Cox accounts of probability

Comparing the Kolmogorov probability axioms eq. (2.74) and the derived properties of the Cox probability assignment eq. (2.78) three observations immediately comes to mind regarding p_K and p_C: (i) p_K is defined on sets in a σ-algebra while p_C is defined on a (product of) Boolean algebras (ii) p_K is a function of one argument while p_C is a function of two arguments (unconditional/conditional probabilities are taken as the basic building block) and (iii) p_K obeys *countable* additivity eq. (2.74c) while p_C obeys *finite* additivity eq. (2.78c). Notice p_F is similar to p_C in these respects and need not be treated explicitly.

Item (i): The spaces p_K and p_C is defined on While this may seem to be a major difference between p_K and p_C in practice it will not matter too much. Firstly, the propositions of practical interests in machine learning will be about numbers and in particular take a form encountered in eq. (2.71). For instance if we consider the value of a real variable x we can consider a set of propositions $A_1, A_2, \ldots$ where

$$A_i = x \text{ is contained in }]a_i, a_{i+1}]. \tag{2.81}$$

Secondly, one will in practice have to be careful when specifying the available collection of sets which can be used to form propositions such as eq. (2.81) since if we allowed propositions of the form $x \in E$ for *all* $E \subset \mathcal{X}$ then this would involve the same difficulties which lead Kolmogorov to define p_K on a σ-algebra. More generally, there are the well-known paradoxes of formal logic such as Russell's paradox [Russell, 1903] which prohibits the use of universal sets such as the set of all true/false propositions. Thirdly, there is Stones theorem [Stone, 1936] which shows a Boolean algebra is isomorphic to a σ-algebra defined in an appropriate space, however we will omit the details here. Taken together, the use of a Boolean algebra of propositions $\mathcal{A}$ or a σ-algebra $\mathfrak{F}$ is not an important distinction from a formal perspective and from a practical perspective, when working with numbers, one would invariably end up with effectively using σ-algebras.

Item (ii): Conditional probability vs. absolute probabilities Kolmogorov probability is a function of a single argument whereas the Cox probabilities are always conditional. This too does not appear to impose a severe

restriction on Kolomogorov probabilities since conditional probabilities can be represented by an appropriately choosen Kolmogorov probability. To take the most basic example, assume $\mathcal{X}$ is discrete, $A, B \in \mathfrak{F}$ and $p_K(B) > 0$ we can define:

$$p_K(A|B) \equiv \frac{p_K(A \cap B)}{p_K(B)}. \tag{2.82}$$

However the general treatment of conditioning, especially in the Bayesian non-parametrics litterature, is a difficult subject which will not be discussed here. A throughout introduction can be found in Orbanz [2012, appendix C]. The upshot is that while having conditional probabilities as the fundamental building block in p_C, this does not to our knowledge limit Kolmogorov probabilities since the most involved examples of conditioning is done within the Kolmogorov framework (see below).

Item (iii): Finite and infinite additivity The most important distinction between the de Finetti/Cox/Jaynes approaches to probabilities and that offered by Kolmogorov is the distinction between finite/infinite additivity (compare eq. (2.74c) to eq. (2.78c)). Jaynes [2003] takes the pragmatic approach of only working with infinite sets when they can be seen as a well-defined limit of finite sets, for instance the Poisson distribution can be seen as a limit of the biomial distribution, however in modern non-parametrics when considering prior processes on infinite-dimensional objects such as probabilities, functions and measures infinite additivity play an indispensable role and an approach to probability which only allows finite additivity is simply insufficient.

2.3.6 Discussion

The Cox/Jaynes approach to probabilities provides a compelling motivation for the use of probabilities to quantify degrees of belief and the result is an elegant and quantitative framework for reasoning about uncertain propositions. The original proof however suffers from certain limitations. Firstly, for general spaces $\mathcal{X}$ the proof requires extra, unnatural conditions such as that of Paris [1994] definition 2.3.4 to be formally correct and in this case $\mathcal{X}$ is *prevented* from being finite. When applying probability theory to truly large spaces, such as all those of interest in modern non-parametrics, infinite additivity is indispensable and must be assumed at some point. For this reason the Kolmogorov account of probabilities should at this point be considered the more fundamental and the more serious treatments of probability theory and non-parametrics such as Kallenberg [2002, 2005] has measure theory at its center and are thus firmly

within the Kolmogorov approach. Textbook treatments of probability which attempts a de Finetti or Cox-type motivation must also assume infinite additivity at some point. For instance the textbook *"Bayesian Theory"* of Bernardo and Smith [2000] starts out with a very throughout discussion of a de Finetti inspired decision-oriented approach to probabilities, then assumes infinite additivity and proceeds within standard measure theory and the Kolmogorov framework for the main technical content of the book. *"Bayesian Data Analysis"* of Gelman, Carlin, Stern, and Rubin [2014] discusses several foundations of probabilities but takes a pragmatic approach:

> *Rather than engage in philosophical debates about the foundations of statistics, however, we prefer to concentrate on the pragmatic advantages of the Bayesian framework, whose flexibility and generality allow it to cope with very complex problems.*
> [Gelman et al., 2014, p. 4]

In the remainder of this thesis we too will assume a standard measure-theoretical foundation of probability theory.

One should not be to dissuaded by these issues. The degree-of-belief interpretation of probabilities is most natural when formulated on a large, fine-grained set of natural-language type propositions (such as the three examples in section 2.2), and it is arguably in this setting the philosophical questions of what probabilities reflects and are to be interpreted is the more relevant. In this case the Paris requirement definition 2.3.4 is not unreasonable and it will be possible to derive p_C and thus provide answers to these questions. A pragmatic person could then feel justified in accepting a degree-of-belief interpretation of probabilities for propositions such as those discussed in section 2.2 and accept the full measure-theoretical account as a mathematical extension with the proviso of being guarded in interpreting the non-parametric results as representing degrees-of-belief and hoping further mathematical developments may discover a Cox-type argument which contains the full measure-theoretical account of probabilities.

Our view is then the rules eq. (2.44) has something important to say about reasoning under uncertainty for propositions which are well-defined (i.e. can potentially be known to be true or false with certainty), and we will in short refer to the interpretation of probabilities as belief, degrees of belief or plausible reasoning as a Bayesian view on probabilities and no longer distinguish between p_K and p_C. An approach to machine learning where these concepts play a central role (or perhaps more simply, where the models are based on manipulating probabilities using eq. (2.44) will be called a Bayesian or probabilistic approach to machine learning.

Given the many alternate ways of handling uncertainty[Zadeh, 1965, 1973, Dempster, 1967, Shafer et al., 1976] it is surprising why a Bayesian approach to uncertainty has played such a central role in machine learning. Potential reasons for this may be pragmatic, that is probability theory is invariantly easier to apply possibly because it is algebraically simple (compared to two-dimensional theories) or that probability theory admits powerful symmetry arguments for *assigning* probabilities (more on this subject in the following chapter).

Alternatively, the reason may be that the data itself suggests uncertainty in the form of probabilities. For instance it may be proposed that input data, on the most natural interpretation, consists of definite logical propositions (e.g. a sensor measures a definite value or a particular test is true or false) and the propositions we typically wish to reason about either *are* binary propositions (for instance unobserved data, in which case they are of the same form as the input data) or alternatively we *want* to treat them as a binary proposition out of scientific habit or to make quantitative statements. For instance, suppose the system should decide if an image contains a dog; we *could* treat this as a fuzzy proposition (the *dog-ness* of the dog), however the actual images will be labelled either as dog or not a dog, and so any fuzzyness will have to be both introduced and removed at intermediate stages of the analysis.

This is however partly a pragmatic concern and reasoning under uncertainty in history, trials or every-day life is reasoning about the truth of propositions formulated in every-day language. Linguistic is often thought to be best analysed in terms of non-classical logic [Zadeh, 1975]. One should therefore be careful not assume the success of Bayesian methods as a tool for handling uncertainty in machine learning naturally translates into an argument in favor of treating all uncertainty using Bayesian methods, let alone human reasoning in general.

2.4 Probabilistic methods in machine learning

We have so far relied on somewhat stringent arguments when deriving and applying the rules of probability theory. However we will now simply assume the reader is familiar with elementary probability and measure theory and feel confident the previous arguments apply in this setting too. A general references to the use of probabilities is Pitman [1993] and references focusing on an probabilistic methods for machine learning are Gelman et al. [2014] and Bernardo and Smith [2000]. A reader interested in advanced references from the perspective of measure theory include Rosenthal [2006] and Kallenberg [2002]. To introduce

standard notation which will be used later for a real variable x we use

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{2.83}$$

to denote the variable is drawn from a normal distribution with density and probability distribution

$$p(x|\mu, \sigma^2) \equiv \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{2.84}$$

$$p(dx|\mu, \sigma^2) \equiv p(x|\mu, \sigma^2)\mu(dx) \tag{2.85}$$

where dx is a small region around x and μ is the Borel measure. When the distinction is important we will typically use large P for the probability distribution.

2.4.1 Models

The previous examples all contained a loose division between what was assumed known (such as it was untrue the man had no girl born on a Tuesday, the sequence of lottery numbers reported in TV and so on) and the quantities of interest (if the man had two girls or the actual lottery sequence drawn). Loosely speaking the variables assumed to have fixed values will be denoted data and the variables we wish to compute probabilities of are denoted the parameters, however we stress this is only a convention. In this language the computation one is often interested in performing takes the form

$$p(\theta|y, \Omega) = \frac{p(y|\theta, \Omega)p(\theta|\Omega)}{p(y|\Omega)} = \frac{p(y|\theta, \Omega)p(\theta|\Omega)}{\int d\theta' \; p(y|\theta', \Omega)p(\theta'|\Omega)}. \tag{2.86}$$

Where data has been denoted by y and parameters by θ. For convenience Ω will often be suppressed. How θ relates to y, that is the specifics of the joint distribution $p(y, \theta|\Omega)$ in eq. (2.86) will be denoted the *model*. $p(\theta|\Omega)$ is often denoted the *prior* and $p(y|\theta, \Omega)$ the *likelihood*. The following example illustrates these definitions.

2.4.2 A simple network model

Consider a $n \times n$ matrix A. If element ij is denoted A_{ij}, the matrix may be taken as representing a network of n vertices such that there is a edge between vertex i and j iff. $A_{ij} = 1$. We will assume the network is symmetric and contains no self-edges, that is, $A_{ij} = A_{ji}$ and $A_{ii} = 0$.

Next we should consider a model for the network, i.e. how A is related to a some parameters θ: $p(A, \theta|\Omega)$. There is no unique answer to how this distribution should be defined and so one is left with an enormous literature on network modelling, some of which we will discuss in chapter 6. Recall in the past examples we had a clearly defined hypothesis we wished to examine (D_{s_0}, J and so on), however in most applications we are only given the data and need to specify a hypothesis. A common strategy for finding a good hypothesis is to consider either how the data may have arisen from a physical process or how one might plausibly describe the data. For networks, the first option is often not feasible and so most literature takes the descriptive approach. One benefit of this approach is if the model is motivated as a description of the data it can lead to results which are easily interpretable.

Consider as an example a simple social network of friendships in a school. In this case the n vertices $i = 1, \ldots, n$ corresponds to pupils and there is an edge between two pupils i and j if they are reported as friends, $A_{ij} = 1$. An intuitive way to describe a friendship network might be as a collection of groups (or communities) of children such that two children in the same community are more likely to be friends than two children in different communities. For instance it might be the boys and girls often report same-sex friendships or, for a larger network, children within the same school or institution are more likely to be friends than children from different schools or institutions.

Suppose there are K groups labelled labelled $1, \ldots, K$ and denote by $z_i \in \{1, \ldots, K\}$ the group child i belongs to. A simple assumption is the probability of friendships is determined by the groups the children belong to and nothing else. For instance if child i belong to group k and child j to group ℓ the probability of a friendship between i and j can be assumed to be constant $\theta_{k\ell}$. In this case

$$p(A_{ij}|z_i = k, z_j = \ell) = \theta_{k\ell} \tag{2.87}$$

Using θ as shorthand for $(\theta_{k\ell})_{k\leq\ell}$ and $z = (z_i)_{i=1}^n$ the assignment of all children to groups we obtain by the product rule:

$$p(A, \theta, z) = p(A|\theta, z)p(\theta|z)p(z) \tag{2.88}$$

Next we turn our attention to $p(\theta|z)$. The most convenient choice is to assume the collection $\theta_{k\ell}$ is iid. and each follow a Beta distributed. For z we can assume there are K groups and each child is assigned to a group independently of the

rest. The various terms become

$$p(A|\theta, z) = \prod_{1\leq i<j\leq n} \theta_{z_i z_j}^{A_{ij}} (1-\theta_{z_i z_j})^{1-A_{ij}} \tag{2.89a}$$

$$p(\theta|z) = \prod_{1\leq k\leq \ell\leq K} \frac{\Gamma(b_1)\Gamma(b_2)}{\Gamma(b_1+b_2)} \theta_{k\ell}^{b_1-1}(1-\theta_{k\ell})^{b_2-1} \tag{2.89b}$$

$$p(z) = \frac{1}{K^n} \tag{2.89c}$$

where $b_1, b_2 > 0$ and K are assumed to take fixed values. Such parameters are typically called *hyperparameters*. A condensed way to describe the model is how it might be used to construct a network randomly (models which easily admit such a description are called *generative* models). The following equations are equivalent to eq. (2.89)

$$\text{for } i = 1, \ldots, n \qquad z_i \qquad \sim \text{Categorical}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right) \tag{2.90a}$$

$$\text{for } 1 \leq k \leq \ell \leq K \qquad \theta_{k\ell}|z \qquad \sim \text{Beta}\,(b_1, b_2) \tag{2.90b}$$

$$\text{for } 1 \leq i < j \leq n \qquad A_{ij}|\theta, z \sim \text{Bernoulli}\left(\theta_{z_i z_j}\right). \tag{2.90c}$$

This type of model is commonly known as a Stochastic Block Model (SBM) (see [White et al., 1976, Holland et al., 1983, Wasserman and Anderson, 1987]) and most of the work in this thesis will revolve around issues easily motivated from the SBM. For instance the above model suffer from some limitations such as the fixed choice of K. This limitation can potentially be overcome in a number of different ways, for instance by flipping a coin until it come up heads (say b flips) then choose K as the b'th prime, however much work in the past decade on Bayesian methods in machine learning has focused on applying non-parametric methods from probability theory which attempt to address this question in a more general manner, and we will describe some of these in chapter 4. Next there is the issue on the particular form of the network model, we will discuss some alternatives in chapter 6. Having chosen a model one can easily assign a probability to each partition z using the sum and product rules

$$p(z|A) = \frac{\int d\theta\; p(z, \theta, A)}{p(A)}. \tag{2.91}$$

While this provides an analytical expression for what we are interested in computing it entails difficulties both in how to represent this distribution (the number of possible partitions is very large) and, in general, how to carry out the integral over the parameters θ. These issues are addressed using sampling schemes (in particular *Markov chain Monte Carlo*) which will be discussed in chapter 5. It might seem at the present point Bayesian methods is about figuring out a particular model (by which we simply mean joint probability density) and applying eq. (2.86), however as we have already seen in section 2.2.1 the problem

of *arriving* at beliefs and the logical consistency requirements eq. (2.44a) and eq. (2.44b) are *not* equivalent, in particular it required the additional desiderata (IIIb) and (IIIc). Since the major goal of a Bayesian approach to machine learning can be said to be about arriving at beliefs this point deserves some attention and this will be the subject of the next chapter.

Chapter 3

Assigning Beliefs

Any complex situation in science or practical life involve forming, assessing or updating beliefs of future or past events, causal mechanisms, intentions and other complex hypothesis. If we restrict ourselves to situations where we can accept the analysis of the past chapter the problems all involve, in one form or another, the assignment of beliefs to appropriately formulated propositions.

The mathematical formalism of probability theory (the consistency requirements eq. (2.44a) and eq. (2.44b)) allow us to analyze the *relationship* between states of beliefs, however it cannot *in itself* be used to *assign* beliefs (see section 2.2.1). This is necessarily so, for the theory aim to describe *all possible* states of belief allowed by the desiderata *and so this class of possible beliefs should be as broad as possible*. Put in a different way, the theory allow us to express the degree of belief of some propositions in light of our degree of beliefs in other propositions, however this relationship should hold regardless of what the degree of belief of those other propositions happen to be and the theory must be able to accommodate this flexibility.

Certainly the consistency requirements eq. (2.44a) and eq. (2.44b) may be used to rule out some states of belief as *incompatible* with each other, say, to believe

about three propositions A, B, C

$$p(AB|C) = 0.9 \tag{3.1}$$
$$p(A|C) = p(B|C) = 0.1 \tag{3.2}$$

however this will not tell us how likely we *should* believe the propositions are.

3.1 The maximum entropy principle

The problem of assigning numerical values to beliefs may be treated in one of two ways. The first way is to treat the problem of arriving at beliefs as an issue which must be resolved pragmatically and on a case-by-case basis; the analysis of Jesus and the lottery in section 2.2.3 is an example of an application of probability theory without definite initial states of belief, as is an application of Bayes theorem in machine learning where priors are assigned based on purely pragmatic consideration. One can then either hope large amounts of data will drown out any larger effects on the result or alternatively interpret the results qualitatively.

The second approach to the problem is to search for a general theory of assigning beliefs. The treatment of mutually exclusive propositions exemplified by the dice in section 2.2.1 was one such example and as we noted it required the additional desiderata (IIIb) and (IIIc), not necessary for arriving at the consistency requirements eq. (2.44). In this chapter we will persue this later program and attempt to derive a more general framework for assigning numerical values to probability than we have seen so far.

Historically, an important leap far beyond the analysis of the dice was taken in 1957 in two seminal publications by E.T. Jaynes [Jaynes, 1957a,b] (see also Jaynes [2003]). Jaynes provided a rule, *the maximum entropy principle* (MEP), for assigning numerical value to probabilities when only partial information is available. This rule is fundamentally tied into Jaynes account of probabilities as beliefs (the same which we have followed in chapter 2) and his original work sought to re-interpret statistical physics as a form of inference where what is being inferred are states of belief about a physical system.

Before discussing the proposal it should be said this program has been a source of controversy in parts of the physics community. Jaynes himself give a personal account of the controversy and objections in Jaynes [1978]; other objections on the use of maximum entropy in statistical physics is given by Penrose [1979], Dougherty [1993], Buck and Macaulay [1991], D'Agostini [1999].

3.1.1 Arriving at beliefs in machine learning

The MEP is not only controversial in the context of statistical physics but has also been the subject to critical discussion and controversy when applied as a principle of inference, see for instance Shimony [1985b], Cardoso Dias and Shimony [1981], Van Fraassen [1981], Van Fraassen et al. [1986]. It can be taken for granted *some* method of assigning numerical values to beliefs must be admitted, at the very least one that admit the "obvious" result of the dice in section 2.2.1. Furthermore the many successful applications of the MEP within statistical mechanics indicate its potency for analyzing assignment numerical values to degrees of belief under partial information [Jaynes, 2003, 1978, Caticha, 2008].

In this section, rather than focusing on the objections, we will give a positive treatment of the problem of arriving at beliefs in a format loosely following that of the previous chapter.

For a reader familiar with the MEP and statistical physics it should be noted the scope of this section is greater than how the MEP is usually applied in machine learning. We are interested in the MEP as *the unique method of statistical inference fulfilling certain consistency requirements*. To put this concretely, consider a standard machine-learning task where data is denoted by $x \in \mathcal{X}$, parameters by $\theta \in \Theta$ and background-information Ω. In slightly unfamiliar terms, our "job" as practitioners of machine learning is to arrive at beliefs over $\theta \in \Theta$, that is, a distribution (deliberately not written as conditional on x) $p_{\text{post}}(\theta)$ which may or may not depend on x and Ω.

In a typical Bayesian setting the solution to this problem consist of two steps:

(i) First, come up with a model $p(x, \theta|\Omega) = p(x|\theta, \Omega)p(\theta|\Omega)$. This task includes defining the space Θ.

(ii) Secondly, apply Bayes theorem to get

$$p(\theta|x, \Omega) = \frac{p(x|\theta, \Omega)p(\theta|\Omega)}{p(x|\Omega)} \tag{3.3}$$

and *consider the left-hand side of Bayes theorem as the beliefs over various parameters* $\theta \in \Theta$ (i.e. $p_{\text{post}}(\theta)$ in the previous formulation).

The MEP will typically enter as part of step (i) when assigning the prior $p(\theta|\Omega)$. This interpretation of Bayes theorem as the *only* tool for arriving at beliefs is however *not* implied by eq. (2.44) and the chapter will rather propose the MEP

as a general tool for inference that (in some cases) will imply the two-step procedure above and sometimes something else.

This view of the MEP as a tool of inference can already be found in Jaynes [1957a,b], however it was first given an axiomatic approach by Shore and Johnson [1980], see also Skilling [1988, 1989] as well as Uffink [1995] for a discussion on these past approaches and their relationship.

The derivation we will present here, along with interpretation and discussion, will in the main follow the work of Caticha and Giffin [2006] which inspired this chapter, though the reader should be aware there are some deviations in section 3.3.3.1. It should be noted the exact status of the present theory and interpretation is still (at least to our knowledge) controversial and especially uniqueness (see the following section) has been the subject of flawed proofs [Shore and Johnson, 1980, Tikochinsky et al., 1984b,a, 1985] as pointed out by Shimony [1985a], Johnson and Shore [1985], Uffink [1995]. A simple proof for uniqueness is also claimed in the appendix of Uffink [1995], however this proof too seems to contain unclear points.

Finally, we would like to draw attention to the excellent book-length notes of Caticha [2008] containing a much longer discussion on the present theory and its interpretation.

3.2 Formulating the problem

What we seek is a tool for arriving at beliefs. As for the derivation of Bayes theorem we will attempt to take an axiomatic approach: We will first clarify the meaning of the statement *"arrive at beliefs"*, then we will arrive at certain desiderata such a process must fulfill and show how they define a unique procedure.

Firstly, to talk of arriving at beliefs implies most of the theory in chapter 2. For simplicity we denote by x all variables of interest and $\mathcal{X}$ the set of all valid settings of x. That is, if we are considering a set of n propositions, $A_1, \ldots, A_n$, x will be a n dimensional binary vectors and in the more familiar situation of data and parameters outlined in eq. (3.3) x will stand for both data and parameter vectors. Accordingly what we are interested in is a distribution p over the space $\mathcal{X}$.

Secondly, the statement "to arrive" implies change. In general terms, we consider a situation where we go from some past state to a new state p. Consider what

the past state may consist of: As a minimum, suppose we have access to a past state of belief q of $\mathcal{X}$ based on careful deliberation. In this case it seems we should at least admit the state of belief implied by q into the past state. How if we really have no informed past beliefs? In the case where $\mathcal{X}$ is discrete and finite, the derivation section 2.2.1 holds and we can conclude $q(x) = \frac{1}{|\mathcal{X}|}$. If $\mathcal{X}$ is not finite, or even worse, if it is continuous $\mathcal{X} = \mathbb{R}^d$, we could loosely argue if $\mathcal{X}$ is divided into non-overlapping sets of equal size an uninformed assignment of belief should not give any preference for any subset implying $q(x) \propto 1$, and this suggest we should always include a prior state of belief q in our background knowledge. Other related arguments on the meaning of ignorance is discussed by Jaynes [2003], and we will so always assume our background knowledge include some past state of belief q.

Thirdly, there must be something which restricts how we change our beliefs, that is, affect our choice of p given the past state q. This information could potentially come in many forms. For instance some states of belief p may lead to decisions which are more costly than other and this would induce a degree of preference amongst various p's, however it is difficult to formalize this in general. One simple situation we must be able to accommodate is if we are told something for certain about x, for instance the value of a particular coordinate. Then our new state of belief p *cannot* reasonably be one which expresses uncertainty about this coordinate or that the coordinate takes a different value than was observed.

Inspired by this example we will limit ourselves to the simplest situation, namely where some p's can be ruled out explicitly or, put in another way, we are only interested in probability assignments $p \in \mathcal{C}$ where $\mathcal{C}$ is the set of allowed assignments. The formulation of the problem at this stage is now [Shore and Johnson, 1980]

> *Given a past state of beliefs q and a set of possible beliefs $\mathcal{C}$ (possibly not containing q) how do we select select which $p \in \mathcal{C}$ best represent our new state of belief?* (3.4)

Or put more simply, how do we update our beliefs from q to p given p must lie in $\mathcal{C}$.

The problem will be solved by posing certain desiderata this update operation must fulfill. Some of these desiderata are similar in spirit to desiderata (I), (II), (IIIa),(IIIb),(IIIc) considered in chapter 2 and which we naturally must insist holds due to our assumptions beliefs are represented by probabilities, however since the new desiderata refers to an update operation and not a set of beliefs their exact meaning will differ and we will therefore state them anew.

3.2.1 Desiderata for a method for updating beliefs

We will assume the following desiderata for the update operation eq. (3.4). While the past references disagree slightly in the formulation, the desiderata were first stated in their the present formulation in [Shore and Johnson, 1980] (notice however application of the desiderate (C3) is controversial[Uffink, 1995]). Later references that make use of the desiderate in their present (or nearly present) form are Uffink [1996], Caticha and Giffin [2006], Caticha [2008]. Notice the requirements have been stated (or partially stated) several times in other contexts. For instance coordinate invariance (desiderata (C2)) form the basis for Jeffreys method of assigning priors [Jeffreys, 1946, Jaynes, 2003]; arguments for the use of entropy as the only measure of uncertainty (and so as a unique way to quantify uncertainty) may be found in Khinchin [1957], Faddeev [1956]; these arguments echo aspects of the seminal work of C. Shannon in 1948 [Shannon, 1948].

More related to the result taken here, and the axiomatic approach of Shore and Johnson [1980], are axiomatic approaches to the principle of minimum cross-entropy which can be found in Fotheringham and O'Kelly [1989], Kannappan [1972] in the discrete case and in Johnson [1979] for the continuous case which is also what we will consider.

With these comments the desiderata considered are [Shore and Johnson, 1980, Caticha and Giffin, 2006, Caticha, 2008]:

(C1) **Uniqueness:** *The update procedure should yield unique results.* (3.5)

That is, the choice of p based on q and $\mathcal{C}$ should be unique. This desiderata can be compared to the desiderata beliefs corresponds to a single number.

(C2) **Invariance:** *The choice of coordinate system should not matter.* (3.6)

For discrete $\mathcal{X}$, this requirement is simply invariance under relabelling. If we assume the densities q and p are well-behaved and $\mathcal{X}$ is not discrete, it is the requirement a smooth bijective change of coordinates $\Gamma : \mathcal{X} \to \mathcal{X}$ should not affect which conclusions we arrive at, see also the approach of Jeffreys [1946] to assigning prior distributions. This desiderata naturally echos desiderata (IIIb) and (IIIc) used for the dice.

(C3) **System independence:** *If two systems are known to be independent, and one receive independent information about them, it should not matter if one treat them separately or jointly.* (3.7)

This desiderata consider the case of two systems a, b such that $\mathcal{X} = \mathcal{X}_a \times \mathcal{X}_b$, $q(x_a, x_b) = q_a(x_a)q_b(x_b)$ and the constraints defining the set $\mathcal{C}$ apply to each of the two systems a, b independently. In this case the desiderata states we may update each system independently or jointly and the two results must agree. A very important point we will return to later is that we *know* the systems to be independent and how this is intepreted quantitatively. It was this interpretation which was a matter of controversy in the work of Shore and Johnson [1980] and we will here follow the interpretation given by Uffink [1996], Caticha and Giffin [2006].

(C4) **Subset independence:** *It should not matter whether one treats disjoint subsets of system states in terms of separate conditional densities or in terms of the full density.* (3.8)

This locality requirement state (for instance) if we update q to p_1 based on some constraint $\mathcal{C}$, then if we resieve additional information *which does not affect a subset* $\mathcal{S} \subseteq \mathcal{X}$ then updating q to p_2 while taking into account both constraints should yeild equal results on $\mathcal{S}$: $p_1(x|x \in \mathcal{S}) = p_2(x|x \in \mathcal{S})$.

Finally there is the principle of minimal update [Caticha and Giffin, 2006]

(C5) **Minimality:** *Beliefs should be updated only to the extend required by new information.* (3.9)

This is for instance saying if q is admissable, that is, $q \in \mathcal{C}$, we should not update our beliefs, $p = q$.

3.3 Derivation of the maximum entropy principle

We will make the assumption not only to solve the problem of *best* selecting p from q and $\mathcal{C}$, but rather that we will consider the problem of finding a functional S (which takes value in the real numbers) which induce an *order of preference* for all $p_1, p_2 \in \mathcal{C}$ through

$$S[p_1, q] \geq S[p_2, q] \text{ implies } p_1 \text{ is at least as preferable as } p_2. \tag{3.10}$$

In this case the optimal $p \in \mathcal{C}$ can be selected[1] as that which is most preferred [Shore and Johnson, 1980, Caticha, 2008]

$$p = \arg\max_{p' \in \mathcal{C}} S[p', q]. \tag{3.11}$$

[1] Assuming $\mathcal{C}$ is closed; we will not treat this and similar technical difficulties here.

The next sections will consist of applying the desiderata (C1)–(C5) to the functional S in order to derive an algebraic expression. This is the same approach taken by Shore and Johnson [1980], however our derivation will follow that of Caticha and Giffin [2006] more closely. We emphasize the arguments are taken from this reference unless otherwise stated and our deviations from their proof are very minor. A reader only interested in the application may skip to eq. (3.85) and section 3.3.5.

3.3.1 Implications of Locality

Desiderata (C4), *subset independence* or *locality*, is understood intuitively as the following restriction: Suppose the set of states $\mathcal{X}$ is decomposed into non-overlapping regions $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ and we impose the restriction the belief the system is in states $x \in X_1$ is constant, for instance $\int_{x\in\mathcal{X}_1} dx\ p(x) = r \in [0, 1]$, then imposing *additional* restrictions on p which *only* depend on the value p take in $\mathcal{X}_2$ will not affect the most preferred value of p in $\mathcal{X}_1$.

This requirement may seem to technical to be accepted at face value so it will be illustrated with an example where $\mathcal{X}$ is discrete. Consider

$$\begin{aligned}\mathcal{X}_1 &= \textit{"Mammals"} = (\textit{"Cat"}, \textit{"Dog"}, \textit{"Cow"})\\ \mathcal{X}_2 &= \textit{"Birds"} = (\textit{"Eagle"}, \textit{"Chicken"}, \textit{"Hummingbird"}).\end{aligned}$$

Suppose we have some prior belief of the above 6 classes of animals, $q(x)$, and based on this we arrived at a certain belief of the animal in question, $p(x)$ (for instance $p(x) = \frac{1}{6}$ for all x). If we are told the belief the animal was one of the *"Mammals"* was correct and should not change, $p(\mathcal{X}_1) = \frac{1}{2}$, however we should take into account the restriction the expected flying altitude of the animal is $100m$:

$$\sum_{x\in X_2} \text{CruisingAltitude}(x) \times p(x) = 100m \tag{3.12}$$

this information *can* for instance make us suppose the animal is more likely to be an eagle than a chicken, however it should *not* make us believe more strongly the animal is a dog than a cow.

Desiderata (C4) decouple the functional S and restrict it to have the form

$$S[p, q] = \int_{x\in\mathcal{X}} dx\ F(p(x), q(x), x) \tag{3.13}$$

for a *function* $F : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{X} \to \mathbb{R}$. The argument follow Caticha and Giffin [2006].

Consider the discrete case $\mathcal{X} = \{x_i\}_{i=1}^n$. In this case S is fully characterized as a function of the $2n$ parameters $p_i \equiv p(x_i), q_i \equiv q(x_i)$:

$$S[p,q] \equiv S(p_1,\ldots,p_n,q_1,\ldots,q_n). \tag{3.14}$$

Assume $\mathcal{X}$ is divided into two non-overlapping parts $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. Since $\mathcal{X}$ is discrete, denote the indexes corresponding to each set by $\mathcal{D}_1$ and $\mathcal{D}_2$:

$$\mathcal{X}_1 = \{x_i\}_{i\in\mathcal{D}_1} \quad \text{and} \quad \mathcal{X}_2 = \{x_i\}_{i\in\mathcal{D}_2}. \tag{3.15}$$

and denote by $p^{(1)} = (p_i)_{i\in\mathcal{D}_1}$ and $p^{(2)} = (p_j)_{j\in\mathcal{D}_2}$ the probabilities corresponding to the two sets. We will use the informal notation $p = (p^{(1)}, p^{(2)})$.

The subset independence desiderata (C4) state any constraint on $\mathcal{X}_2$ will not affect the *conditional* assignment of probabilities in $\mathcal{X}_1$:

$$p(x_i|x_i \in \mathcal{X}_1) = \frac{p_i}{\sum_{i\in\mathcal{D}_1} p_i}. \tag{3.16}$$

Restrictions on $\mathcal{X}_2$ may however affect the numerical value of a $p_i, i \in \mathcal{X}_i$ by an overall multiplicative factor. To deal with this complication we will assume $\sum_{i\in\mathcal{D}_1} p_i = r$ is constant and the restriction imposed on $\mathcal{X}_1, \mathcal{X}_2$ mean $p^{(1)}, p^{(2)}$ are restricted to manifolds parameterized by $u \in \mathbb{R}^{d_1}$, $1 \le d_1 \le |\mathcal{D}_1| - 1$ and $v \in \mathbb{R}^{d_2}$, $1 \le d_2 \le |\mathcal{D}_2| - 1$

$$u \mapsto p^{(1)}(u) = (p_i(u))_{i\in\mathcal{D}_1}, \quad v \mapsto p^{(2)}(v) = (p_j(v))_{j\in\mathcal{D}_2}, \tag{3.17}$$

and such that $\sum_{i\in\mathcal{D}_1} p_i(u) = r$, $\sum_{j\in\mathcal{D}_2} p_j(v) = 1 - r$ are constant for all u, v. The most preferred value of p is then $\hat{p} \equiv (\hat{p}^{(1)}, \hat{p}^{(2)}) \equiv (p^{(1)}(\hat{u}), p^{(2)}(\hat{v}))$ where each coordinate $\hat{u}_k$ $\hat{u}$ fulfill

$$\frac{\partial}{\partial u_k} S[p,q]\Big|_{(u,v)=(\hat{u},\hat{v})} = \sum_{i\in\mathcal{D}_1} \frac{\partial S[(p^{(1)},p^{(2)}),q]}{\partial p_i} \frac{\partial p_i(u)}{\partial u_k}\Big|_{(u,v)=(\hat{u},\hat{v})} = 0. \tag{3.18}$$

Introducing $f_i(p,q) = \frac{\partial S[p,q]}{\partial p_i}$, keeping the parametrization of $p^{(1)}$ fixed and noticing this must hold for *any* restrictions imposed on $\mathcal{X}_2$, particulary to restricting $p^{(2)}$ to take any value that sum to $1-r$, we have by subset independence for each k and $p^{(2)}$ such that $\sum_j p_j^{(2)} = 1 - r$:

$$\sum_{i\in\mathcal{D}_1} f_i(p^{(1)}(\hat{u}), p^{(2)}, q) h_{ik} = 0 \text{ where } h_{ik} = \frac{\partial p_i(u)}{\partial u_k}\Big|_{u=\hat{u}}. \tag{3.19}$$

The only way for this to hold for general linear combinations is if the functions $f_i(p,q), i \in \mathcal{D}_1$ are independent of $p_j, j \in \mathcal{D}_2$ and since the domains $\mathcal{D}_1, \mathcal{D}_2$ are

arbitrary we must have that $\frac{\partial S}{\partial p_i}$ is independent of all $(p_j)_{j\neq i}$ and so eq. (3.14) become

$$\frac{\partial S[p,q]}{\partial p_i} = f_i(p_i, q_1, \ldots, q_n). \tag{3.20}$$

A similar argument can be carried through for q. Consider a small variation of q on $\mathcal{D}_2$. That is a small vector $\delta q^{(2)}$ such that

$$\sum_{j=1}^{n} \delta q_j^{(2)} = 0 \text{ and for } j \in \mathcal{D}_1\text{:} \quad \delta q_j^{(2)} = 0. \tag{3.21}$$

In this case eq. (3.20) become for $i \in \mathcal{D}_1$

$$\frac{\partial S[p,q]}{\partial p_i} = f_i(p_i, q + \delta q^{(2)}). \tag{3.22}$$

However from the subset independence desiderata local changes to $\mathcal{D}_2$ must have no effect on $\mathcal{D}_2$ and comparing to eq. (3.19) for this to have no change on the local maxima of $p^{(1)}$, f_i must be independent of $\delta q^{(2)}$. In this case eq. (3.22) is simply

$$\frac{\partial S[p,q]}{\partial p_i} = f_i(p_i, q_i). \tag{3.23}$$

Integrating we get

$$S[p,q] = \sum_{i=1}^{n} F_i(p_i, q_i) + \tilde{F}(q). \tag{3.24}$$

for functions F_i and $\tilde{F}$. Since $\tilde{F}$ does not affect the choice of p it may be omitted without loss of generality. Taking the continuum limit we arrive at eq. (3.13): $S[p,q] = \int_{x\in\mathcal{X}} dx\ F(p(x), q(x), x)$.

3.3.2 Implications of coordinate invariance

Next we show coordinate invariance (C2) implies eq. (3.13) take the form

$$S[p,q] = \int dx\ q(x)\Phi\left(\frac{p(x)}{q(x)}\right) \tag{3.25}$$

for some function Φ. The argument follow that of Caticha [2008] with some minor variations. Consider maximizing the objective eq. (3.13):

$$S[p,q] = \int_{x\in\mathcal{X}} dx\ F(p(x), q(x), x) \tag{3.26}$$

under a single linear constraint $\int dx\ a(x)p(x) = 0$. By calculus of variation it follows the posterior satisfy (for all x):

$$\lambda + \mu a(x) + F_1(p(x), q(x), x) = 0 \tag{3.27}$$

for Lagrange multipliers λ, μ and using the convention from chapter 2 that F_1 correspond to the derivative of F with respect to the first coordinate. Next, consider any smooth coordinate transformation $\Gamma : \mathcal{X} \to \mathcal{X}$, i.e. $x = \Gamma(y)$. The transformed prior density $q'(y)$ is then

$$q(x) = q'(y) \left| \det \left(\frac{\partial y}{\partial x} \right) \right| = q'(y) J(x) \tag{3.28}$$

where $J(x)$ is the determinant of the Jacobian of the inverse of the mapping Γ evaluated at x. We also assume the transformed constraint function a' fulfill $a'(y) = a'(\Gamma(x)) = a(x)$. In these coordinates and for all y:

$$\lambda' + \mu' a'(y) + F_1(p'(y), q'(y), y) = 0 \tag{3.29}$$

for new Lagrange multipliers λ', μ'. By insertion eq. (3.29) is easily seen to be equivalent to

$$\lambda' + \mu' a(x) + F_1(J^{-1}p(x), J^{-1}q(x), \Gamma(x)) = 0. \tag{3.30}$$

Combining eq. (3.30) and eq. (3.27) we arrive at

$$\frac{F_1(J^{-1}p(x), J^{-1}q(x), \Gamma(x))}{F_1(p(x), q(x), x)} = \frac{\lambda' + a(x)\mu'}{\lambda + a(x)\mu}. \tag{3.31}$$

Next suppose $\mathcal{X} = \cup_{k=1}^{K} \mathcal{X}_k$ is a partitioning of $\mathcal{X}$ and $q(x), a(x)$ is constant in each $\mathcal{X}_k$. Since the system of coordinates carry no information it follows $p(x)$ must be constant too in each $\mathcal{X}_k$. In this case for *any* fixed transformation eq. (3.31) reduces to

$$\text{for all } x \in \mathcal{X}_k\text{: } \frac{F_1(J^{-1}(x)p_k, J^{-1}(x)q_k, \Gamma(x))}{F_1(p_k, q_k, x)} = C_k. \tag{3.32}$$

for constant C_k. Suppose in the region $\mathcal{X}_k$ there are two subsets $A, B \subset \mathcal{X}$ such that the mapping Γ is the identity $\Gamma(x) = x$ for $x \in A$ and has $J(x) = 1$ for $x \in B$ but is *not* the identity. Considering $x \in A$, the left-hand side of eq. (3.32) is 1. Thus $C_k = 1$ within $\mathcal{X}$ and so for $x \in B$:

$$\text{for all } x \in B\text{: } \frac{F_1(p_k, q_k, \Gamma(x))}{F_1(p_k, q_k, x)} = 1. \tag{3.33}$$

For this to hold in the general case F_1 must be independent of x. Next suppose Γ is still the identity on A but an arbitrary smooth transformation on B. In this case

$$\text{for all } x \in B\text{: } \frac{F_1(J^{-1}(x)p_k, J^{-1}(x)q_k)}{F_1(p_k, q_k)} = 1. \tag{3.34}$$

For eq. (3.34) to hold for *arbitrary* transformations and so arbitrary $J^{-1}(x)$ it follows F_1 can only depend on the ratio of it's arguments, i.e. there is a function $\tilde{h}$ such that

$$F_1(x,y) = \tilde{h}\left(\frac{x}{y}\right) \tag{3.35}$$

which implies the general solution

$$F(x,y) = xh\left(\frac{x}{y}\right) + v(y) \tag{3.36}$$

for functions h and v. Since the function v will not affect the minimum it may be omitted. Introducing $\Phi(x) = xh(x)$ we obtain eq. (3.25).

3.3.3 Subsystem independence

The next step is to limit the functional form of Φ in eq. (3.25). We will first show due to desiderata (C3) the available choices of Φ limit S to be equivalent (up to monotone transformations which preserve order such as scaling and translation) to the following Rényi entropy-like functions [Rényi, 1962] parameterized by $\eta > -1$ (in the following we will simply denote this as a Rényi entropy):

$$S[p,q] = U_\eta[p,q] = \frac{1}{\eta(1-\eta)}\left(1 - \int dx\ p(x)\left(\frac{p(x)}{q(x)}\right)^\eta\right). \tag{3.37a}$$

It is instructive to examine the limits $\eta \to 0$ and $\eta \to -1$, the former will be of particular interest later. Taylor expanding in η and making use of $\log(1+\varepsilon) \approx \varepsilon + \mathcal{O}(\varepsilon^2)$ and $x^\varepsilon \approx 1 + \varepsilon \log x + \mathcal{O}(\varepsilon^2)$ we obtain by l'Hôpital's rule for $\eta \to 0$ (see [Hardy et al., 1952])

$$\begin{aligned}\lim_{\eta\to 0} U_\eta[q,p] &= \lim_{\eta\to 0} \frac{1}{1-2\eta}\frac{d}{d\eta}\left(1 - \int dx\ p(x)\left[1 + \eta \log\frac{p(x)}{q(x)}\right]\right) \\ &= -\int dx\ p(x)\log\frac{p(x)}{q(x)}.\end{aligned} \tag{3.38}$$

And for $\eta \to -1$

$$\begin{aligned}\lim_{\eta\to -1} U_\eta[q,p] &= \lim_{\eta\to -1} \frac{1}{\eta(1-\eta)}\left(1 - \int dx\ q(x)\left(\frac{p(x)}{q(x)}\right)^{\eta+1}\right) \\ &= \lim_{\eta\to -1} \frac{1}{1-2\eta}\frac{d}{d\eta}\left(1 - \int dx\ p(x)\left[1 + (\eta+1)\log\frac{p(x)}{q(x)}\right]\right) \\ &= \int dx\ q(x)\log\frac{p(x)}{q(x)}.\end{aligned} \tag{3.39}$$

For other limiting cases see Uffink [1995], Wootters [1981].

Before addressing the general result we need to formalize the notion of desiderata (C3), system independence. Loosely speaking, system independence mean systems which are independent must have independent solutions, that is, if they are treated independently or jointly they will lead us to select the same posterior. The intuition behind this requirement is simple. Suppose there are two systems a, b composed of two blackjack tables, one in Las Vegas and the other at a casino in Alpha-Centauri and suppose we have prior belief q_a and q_b on the state of the tables x_a, x_b which is independent: $q_{ab}(x_a, x_b) = q_a(x_a)q_b(x_b)$. Then if we receive two pieces of independent information, one which only relate to table a and one which only relate to table b, we will draw the same conclusion about the tables if we treat them jointly or independently.

3.3.3.1 Formalization of subsystem independence

There is a subtle point in how to formalize subsystem independence which highlights some important aspects on appropriately processing restrictions raised by Uffink [1995]. In Shore and Johnson [1980] the axiom was given the following interpretation: Consider again two systems a, b with prior densities q_a, q_b and assume we obtain new information in the form of two constraints on a and b, for instance

$$\int dx\ p_a(x)f_a(x) = 1 \qquad \int dx\ p_b(x)f_b(x) = 1 \tag{3.40}$$

Let $\mathcal{C}_a$ and $\mathcal{C}_b$ be the space of distributions satisfying eq. (3.40) and $\mathcal{C}_{ab}$ the space of joint distributions $p_{ab}(x_a, x_b)$ satisfying both requirements marginally, that is if $p_{ab} \in \mathcal{C}_{ab}$: $\int dx_b\ p_{ab}(\cdot, x_b) \in \mathcal{C}_a$ and $\int dx_a\ p_{ab}(x_a, \cdot) \in \mathcal{C}_b$. According to Shore and Johnson [1980] subsystem independence means either processing the two systems independently or separately will result in the same posterior density. Specifically if

$$p_a = \arg\max_{p \in \mathcal{C}_a} S[p, q_a] \tag{3.41a}$$

$$p_b = \arg\max_{p \in \mathcal{C}_b} S[p, q_b] \tag{3.41b}$$

$$p_{ab} = \arg\max_{p \in \mathcal{C}_{ab}} S[p, q_a q_b] \tag{3.41c}$$

then $p_a(x_a)p_b(x_b) = p_{ab}(x_a, x_b)$ for all x_a, x_b.

The key point is the following: In the above treatment we merely assumed our *prior* belief was independent and the information we received referred to the

marginal distribution over the systems a, b, however in this case we would do better by formulating subsystem independence to refer to *that* situation and *not* use the formulation that *"the systems are independent"*.

The problem is that simply because we do not a-priori believe two systems are dependent (that is, our prior belief factorize: $q_{ab} = q_a q_b$), and we obtain independent information, it is much less clear why we should *require* our final state of belief to be independent too. After all, it might be the case the systems *were in fact* dependent and we would certainly not have erred in having a posterior belief with some dependency!

To give an example analogous to Uffink [1995], consider a system composed of two hands a, b and the state of the system refer to the skin color of the hands which can be either black or white. We assume all four states of coloring of the hands have equivalent probability of $\frac{1}{4}$. Now we are given the (independent) pieces of information that *"Hand a (or b) is attached the secretary of the prime minister in Brazil"*. Assuming we do not know anything about the skin color of the secretary of the prime minster in Brazil this tell us nothing about the color of each hand a,b, however we would certainly now suspect the hands have the same color, whatever it is!.

We could criticize the above argument by saying our background information that the same person has the same color of hands would have made our factorized prior impossible, or should have resulted in us considering a more complicated model from the outset. This might be the case, however the problem still remains that the only reason we should admit a desiderata such as subsystem independence is if we feel compelled to do so, and the only reason we would feel truly compelled to do so is if the systems *are actually known to be independent*, e.g. if one hand is known to be in China and the other hand is known to be in France. Following Karbelkar [1986], Uffink [1995] we will therefore assume *"system independence"* refer to some physical or logical property which implies the *posterior* too factorizes:

$$p_{ab}(x_a, x_b) = p_a(x_a) p_b(x_b). \tag{3.42}$$

Specifically, subsystem independence means the following two optimization tasks should give equivalent results

(i) Determine the distribution $p_{ab}(x_a, x_b) = p_a(x_a) p_b(x_b)$ according to

$$\underset{p_a \in \mathcal{C}_a, p_b \in \mathcal{C}_b}{\arg\max} \int dx_a \; q_a(x_a) \Phi\left(\frac{p_a(x_a)}{q_a(x_a)}\right) \int dx_b \; q_b(x_b) \Phi\left(\frac{p_b(x_b)}{q_b(x_b)}\right) \tag{3.43}$$

(ii) Determine the distribution $p_{ab}(x_a, x_b) = p_a(x_a)p_b(x_b)$ according to

$$\underset{p_a \in \mathcal{C}_a, p_b \in \mathcal{C}_b}{\arg\max} \int dx_a dx_b \; q_a(x_a) q_b(x_b) \Phi\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) \tag{3.44}$$

Notice this is a much weaker requirement than that of Shore and Johnson [1980].

3.3.3.2 Consequence of subsystem independence

Having arrived at a formal translation of system independence we are now ready to show the functional eq. (3.27):

$$S[p, q] = \int dx \; q(x) \Phi\left(\frac{p(x)}{q(x)}\right) \tag{3.45}$$

corresponds to a Rényi entropy U_η. Historically, the first proof this was so was by Shore and Johnson [1980], however as argued above this proof assumes a too strong definition of independence, namely in that it uses a formulation of system independence which is stronger than implied by the desiderata. The second proof is due to Uffink [1995], however this proof too seem to contain a difficulty. In [Uffink, 1995, p. 258] it is assumed for a system of 3 states and two densities q_2, p_2 that: $\frac{q_2(x_i)}{p_2(x_i)} = \alpha$ for $i = 1, \ldots, 3$. It follows $\alpha = 1$, however the proof rest on characterizing a linear form by varying α. This is likely due to a typesetting issue or a misunderstanding on our part, however we have been unable to tell which. The final proof is due to Caticha and Giffin [2006] which we follow below, however we will slightly extend the first part of the argument.

Consider first the case where the belief about systems a, b are obtained independently according to method (i), eq. (3.43) under the constraints eq. (3.40). In this case

$$\dot{\Phi}\left(\frac{p_a(x_a)}{q_a(x_a)}\right) = \mu_a f_a(x_a) + \kappa_a \tag{3.46a}$$

$$\dot{\Phi}\left(\frac{p_b(x_b)}{q_b(x_b)}\right) = \mu_b f_b(x_b) + \kappa_b \tag{3.46b}$$

for Lagrange multipliers μ_a, μ_b and κ_a, κ_b. Next we treat the systems jointly according to method (ii), eq. (3.44). In this case the variational problem becomes

$$\delta \Big[S[p_a p_b, q_a q_b] - \alpha \int dx_a dx_b \;\; (p_a(x_a)p_b(x_b) - 1) \\ -\lambda_a \int dx_a \;\; (f_a(x_a)p_a(x_a) - 1) - \lambda_b \int dx_b \;\; (f_b(x_a)p_b(x_b) - 1) \Big]. \tag{3.47}$$

For our purpose we only need to consider the variation with respect to $p_a(x_a)$ and $p_b(x_b)$:

$$D_a(x_a) \equiv \int dx_b\ p_b(x_b)\dot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) \qquad = \lambda_a f_a(x_a) + \alpha \tag{3.48a}$$

$$D_b(x_b) \equiv \int dx_a\ p_a(x_a)\dot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) \qquad = \lambda_b f_b(x_b) + \alpha. \tag{3.48b}$$

Multiplying eq. (3.48) by $p_b(x_b)$ and $p_a(x_a)$ respectively, summing over x_a and x_b and using that the constraints eq. (3.40) normalize to 1 we obtain eq. (3.51) with the definitions eq. (3.49) and eq. (3.50):

$$Q[p_a, p_b] \equiv \int dx_a dx_b\ p_a(x_a)p_b(x_b)\dot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) \tag{3.49}$$

$$\lambda \equiv \lambda_a = \lambda_b. \tag{3.50}$$

$$Q[p_a, p_b] = \lambda + \alpha. \tag{3.51}$$

Using eq. (3.51) and letting Q be shorthand for $Q[p_a, p_b]$ we can eliminate the unknown multipliers in eq. (3.48):

$$D_a(x_a) = \lambda f_a(x_a) + Q - \lambda = \lambda(f_a(x_a) - 1) + Q \tag{3.52a}$$

$$D_b(x_b) = \lambda f_b(x_b) + Q - \lambda\ = \lambda(f_b(x_b) - 1) + Q. \tag{3.52b}$$

Eliminating λ in eq. (3.52) and simplifying gives

$$D_a(x_a) = Q + \frac{(D_b(x_b) - Q)(f_a(x_a) - 1)}{f_b(x_b) - 1} \tag{3.53}$$

Inserting the definition of $D_a(x_a)$ from eq. (3.48a) into eq. (3.53) and re-ordering the terms on the left-hand side we obtain

$$\int dx_b\ p_b(x_b)\dot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) = f_a(x_a)\frac{D_b(x_b) - Q}{f_b(x_b) - 1} + \frac{-D_b(x_b) + Qf_b(x_b)}{f_b(x_b) - 1}. \tag{3.54}$$

Taking the functional derivative with respect to $p_b(x_b)$ on both sides of eq. (3.54) and using the assumption the optimum found using method (i) and (ii) must be equivalent to eliminate $f_a(x_a)$ using eq. (3.46) we obtain:

$$\dot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) + \frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\ddot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) = \dot{\Phi}\left(\frac{p_a(x_a)}{q_a(x_a)}\right)M_b + K_b \tag{3.55}$$

where M_b and K_b are defined as

$$M_b \equiv \frac{1}{\mu_a}\frac{\delta}{\delta p_b(x_b)}\left[\frac{D_b(x_b) - Q}{f_b(x_b) - 1}\right] \tag{3.56}$$

$$K_b \equiv \frac{\delta}{\delta p_b(x_b)}\left[\frac{-D_b(x_b) + Qf_b(x_b)}{f_b(x_b) - 1}\right] - \kappa_a M_b. \tag{3.57}$$

Notice that in the argument of the functional derivatives in the definition of M_b and K_b the parameter x_a is integrated out. Taking the functional derivative with respect to $p_b(x_b)$ does not change that. Accordingly M_b and K_b are *not* functions of x_a. Thus, if we differentiate eq. (3.55) after x_a and solve for M_b we obtain

$$M_b = \left[\frac{d}{dx_a}\dot{\Phi}\left(\frac{p_a(x_a)}{q_a(x_a)}\right)\right]^{-1} \times \left[\frac{d}{dx_a}\left(\dot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right) + \frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\ddot{\Phi}\left(\frac{p_a(x_a)p_b(x_b)}{q_a(x_a)q_b(x_b)}\right)\right)\right]. \tag{3.58}$$

Since the expressions on the right-hand side are functions of $y_a \equiv \frac{p_a(x_a)}{q_a(x_a)}$ and $y_b \equiv \frac{p_b(x_b)}{q_b(x_b)}$ so is M_b. However since

$$0 = \frac{\partial M_b}{\partial x_a} = \frac{\partial y_a(x_a)}{\partial x_a}\left[\frac{\partial M_b(y_a, y_b)}{\partial y_a}\right] \tag{3.59}$$

hold for arbitrary prior beliefs $q_a(x_a)$ it follows $\frac{\partial M_b(y_a,y_b)}{\partial y_a} = 0$. Accordingly we can consider M_b to be a function of y_b alone and write $M_b \equiv M_b(y_b)$; a similar argument show $K_b \equiv K_b(y_b)$ too and considering the system a allows us to define $M_a(y_a)$ and $K_a(y_a)$. Using the symmetry of the left-hand side of eq. (3.55) we arrive at

$$\dot{\Phi}\left(\frac{p_a(x_a)}{q_a(x_a)}\right)M_b(y_b) + K_b(y_b) = \dot{\Phi}\left(\frac{p_b(x_b)}{q_b(x_b)}\right)M_a(y_a) + K_a(y_a). \tag{3.60}$$

From here on the rest of the argument is also found in Caticha and Giffin [2006]. Consider the case where the space of x_a and x_b are equivalent and we have the same prior information $q_a = q_b$ and the same constraints $f_a = f_b$. In this case it follows by eq. (3.46) the optimal values are equal too and we can choose $\mu_a = \mu_b \equiv \mu$ and $\kappa_a = \kappa_b \equiv \kappa$. It also follow by definitions that $M_a(y) = M_b(y) \equiv M(y)$, $K_a(y) = K_b(y) \equiv K(y)$. We can then consider eq. (3.60)

$$\frac{\dot{\Phi}(y_b) - \kappa}{\mu}M(y_a) + K(y_a) = \frac{\dot{\Phi}(y_a) - \kappa}{\mu}M(y_b) + K(y_b). \tag{3.61}$$

Computing the derivative $\frac{\partial}{\partial y_b}$ of eq. (3.61) to obtain eq. (3.62) and $\frac{\partial^2}{\partial y_a \partial y_b}$ to obtain eq. (3.63) we get by rearranging:

$$\ddot{\Phi}(y_b)M(y_a) = \dot{\Phi}(y_a)\dot{M}(y_b) + \dot{K}(y_b) \tag{3.62}$$

$$\frac{\dot{M}(y_a)}{\ddot{\Phi}(y_a)} = \frac{\dot{M}(y_b)}{\ddot{\Phi}(y_b)}. \tag{3.63}$$

Keeping y_b fixed and considering eq. (3.63) a function of y_a we can integrate and obtain:

$$M(y_a) = c_0\dot{\Phi}(y_a) + d_0 \tag{3.64}$$

for constants $c_0 = \frac{\dot{M}(y_b)}{\ddot{\Phi}(y_b)}$ and d_0. Plugging this into eq. (3.62) and using eq. (3.63) we have

$$\begin{aligned}
\dot{K}(y_b) &= \ddot{\Phi}(y_b)\left(c_0\dot{\Phi}(y_a) + d_0\right) - \Phi(y_a)\dot{M}(y_b) \\
&= \ddot{\Phi}(y_b)\left(\frac{\dot{M}(y_b)}{\ddot{\Phi}(y_b)}\dot{\Phi}(y_a) + d_0\right) - \Phi(y_a)\dot{M}(y_b) \\
&= d_0\ddot{\Phi}(y_b) \qquad (3.65) \\
K(y_b) &= d_0\dot{\Phi}(y_b) + f_0 \qquad (3.66)
\end{aligned}$$

for constant f_0. Substituting this into eq. (3.60) we obtain

$$\dot{\Phi}(y_a y_b) + y_a y_b\ddot{\Phi}(y_a y_b) = c_0\dot{\Phi}(y_a)\dot{\Phi}(y_b) + d_0\left[\dot{\Phi}(y_a) + \dot{\Phi}(y_b)\right] + f_0. \tag{3.67}$$

Since this equation must hold for general problems we can consider a particular problem in which $y_b = \frac{p_b(x_b)}{q_b(x_b)}$ take the value 1. Fixing y_b to 1 and considering $y \equiv y_a$ we obtain

$$\dot{\Phi}(y) + y\ddot{\Phi}(y) = c_0\dot{\Phi}(y)\dot{\Phi}(1) + d_0\left[\dot{\Phi}(y) + \dot{\Phi}(1)\right] + f_0. \tag{3.68}$$

Taking the derivative of this equation with respect to y and defining $\eta \equiv c_0\dot{\Phi}(1) + d_0 - 1$ this reduces to

$$y\dddot{\Phi}(y) = (1-\eta)\ddot{\Phi}(y). \tag{3.69}$$

Recall the objective is to determine Φ. It is easy to see eq. (3.69) behaves differently when integrated if $\eta = 0$ or $\eta = -1$ (The case $\eta = 1$ is not interesting and will not be considered). We treat the different possibilities for η below

The case $\eta \neq -1, 0$ Integrating eq. (3.69) twice we finally obtain

$$\dot{\Phi}(y) = uy^\eta + v. \tag{3.70}$$

The above was derived under the assumption $y_b = 1$ and need not hold for all y_b. Indeed, plugging eq. (3.70) into eq. (3.67) result in the necessary conditions

$$\begin{aligned}
u(1+\eta) &= c_0u^2 \qquad (3.71a) \\
0 &= c_0uv + d_0v \qquad (3.71b) \\
v &= c_0v^2 + 2d_0v + c. \qquad (3.71c)
\end{aligned}$$

The solution $u = 0$ result in a order of preference independent of the distribution p and can be ruled out. There remain three non-trivial solutions

$$u = \frac{1+\eta}{c_0} \qquad v = \frac{-d_0}{c_0} \qquad f_0 = \frac{d_0(1-d_0)}{4c_0}. \tag{3.72}$$

Inserting this into eq. (3.70), integrating and inserting the solution of Φ into the definition of $S[p,q]$ in eq. (3.46) we obtain

$$\Phi(y) = \frac{1}{c_0} y^{\eta+1} - \frac{d_0}{c_0} y + C \tag{3.73}$$

$$S[p,q] = U_\eta[p,q] = \frac{1}{c_0} \int dx\ p(x) \left(\frac{p(x)}{q(x)}\right)^\eta - \frac{d_0}{c_0} + C \tag{3.74}$$

which is equal to the desired result up to a linear transformation.

The case $\eta = 0$ If $\eta = 0$ it is easy to verify eq. (3.69) and eq. (3.46) has the solution

$$\Phi(y) = u' y \log y + v' y + w' \tag{3.75}$$

$$S[p,q] = U_0[p,q] = u' \int dx\ p(x) \log \frac{p(x)}{q(x)} + v' + w' \tag{3.76}$$

which is equivalent to a linear transformation of the ordinary conditional entropy.

The case $\eta = -1$ Finally if $\eta = -1$ we arrive at the solution

$$\Phi(y) = u'' \log y + v'' y + w'' \tag{3.77}$$

$$S[p,q] = U_{-1}[p,q] = u'' \int dx\ q(x) \log \frac{p(x)}{q(x)} + v'' + w'' \tag{3.78}$$

which is equivalent to a linear transformation of the (reverse) conditional entropy.

3.3.4 Selecting the right entropy

Having shown the desiderata naturally point to the order of preference $S[p,q]$ as being a Rényi entropy-like function U_η characterized by η it is natural to ask which value of η is the correct one. This endeavor face both hindrance and encouragement. The hindrance is the desiderata considered so far *cannot*

fix η since it can be shown all possible values of η will satisfy them (see for instance Uffink [1995], Karbelkar [1986]). Fortunately there is also good news. If the desiderata are only satisfied by a single type of functional (up to order-preserving transformations) characterized by a single parameter we are able to treat the problem empirically.

In our case there is overwhelming physical evidence that $\eta = 0$ from thermodynamics Jaynes [1957a,b, 1989, 2003] which force us to either accept $\eta \approx 0$, or if we suggest a value of η very different from 0, then we must either showing how these applications are defective or show some physical considerations exempt these cases from one of the considered desiderata. For completeness, we will briefly sketch one such argument based on the weak law of large numbers and subsystem independence (desiderata (C3)). The argument was first considered by I. Csiszár in [Bernardo, 1985, p. 83] (see also Grendár Jr and Grendár [2003]), however our presentation will follow that of Caticha and Giffin [2006], Caticha [2008].

Consider a set of n systems and assume each system can be in m distinct states $i = 1, \ldots, m$. Each system is given identical priors $q_i \equiv q(i)$ and the posterior distributions, p_i, are each subject to a linear constraint such as

$$\sum_{i=1}^{m} p_i f_i = E. \tag{3.79}$$

A typical example of this situation is a finite Ising spin lattices with k spins [Onsager, 1944, Lenz, 1920] where (assuming no symmetries) $m = 2^k$ and the above constraint would correspond to the energy. The system can now be treated in two ways.

Independent treatment Assume η is at it's correct value and letting $\mathcal{C}_F$ be all posteriors fulfilling eq. (3.79), the correct (most preferred) posteriors can then be found by maximizing the order of preference

$$p_{\text{opt}} = \arg\max_{p \in \mathcal{C}_F} U_\eta[p, q] \tag{3.80}$$

Joint treatment We could also imagine preparing a large number of equivalent spin systems and allowing them to thermalize by bringing them into contact with a heatbath with the same temperature as the single system. If we measure the state of each system at any given point, we expect (in the limit of many systems) that the number of systems found in any given state i, n_i, should be proportional to the frequency we expect from the single system case and the

average energy found in the n systems is equal to the energy constraint in the single-system case:

$$\tilde{p}_i \equiv \frac{n_i}{n} \tag{3.81}$$

$$\tilde{E} \equiv \sum_{i=1}^{m} \tilde{p}_i f_i \tag{3.82}$$

and $\tilde{p}_i \to p_i, \tilde{E} \to E$ for $n \to \infty$. Ignoring the equality constraint eq. (3.79) and only considering the prior distribution, the probability of a single observation of frequencies $\tilde{p} = (\frac{n_i}{n})_i$ is given by the multinomial distribution

$$p_n(\tilde{p}|q) = \frac{n!}{\prod_{i=1}^{m} n_i!} \prod_{i=1}^{m} q_i^{n_i}, \qquad \sum_{i=1}^{m} n_i = n. \tag{3.83}$$

Using Stirlings approximation $\log n! = n \log n - n + \log \sqrt{2\pi n} + \mathcal{O}(\frac{1}{n})$ we obtain

$$\begin{aligned} \frac{1}{n} \log p_n(\tilde{p}|q) &\approx \sum_{i=1}^{m} \frac{n_i}{n} \log \frac{n_i}{n q_i} - \sum_{i=1}^{m} \log \sqrt{\frac{n_i}{n}} - (n-1) \log \sqrt{2\pi n} \\ &= U_{\eta=0}[\tilde{p}; q] - \sum_{i=1}^{m} \frac{1}{2n} \log \tilde{p}_i - \frac{n-1}{n} \log \sqrt{2\pi n}. \end{aligned} \tag{3.84}$$

In the large-n limit the entropy-term will dominate, and deviations from the true frequency will be more and more disfavored. If we introduce Lagrange multipliers to satisfy the energy constraint eq. (3.79) *for the entire ensemble* the argument becomes slightly more complicated because the values of E that can be realized will depend on n, however intuitively the most preferred distribution in the joint case is obtained by maximizing the $\eta = 0$ entropy subject to a linear constraint in a similar fashion as Jaynes [1957a]. A more rigorous treatment of these issues is given by I. Csiszár in [Bernardo, 1985, p. 83].

Since per assumption the joint and independent treatment should be equivalent we can conclude η must be equivalent for the joint and independent treatment and thus we arrive at the familiar result for the order of preference S [Caticha and Giffin, 2006, Caticha, 2008].

$$S[p, q] = U_{\eta=0} = -\int dx\ p(x) \log \frac{p(x)}{q(x)} \tag{3.85}$$

3.3.5 Bayes rule as a special case of ME

Consider again the situation outline in section 3.1.1 where we are given a data vector $x \in \mathcal{X}$ and we wish to use an observation of x to draw conclusions about

a vector of parameters $\theta \in \Theta$. An examples could be where x corresponds to a sequence of flips of a coin and θ corresponds to the probability the next flip is heads. Suppose we observe a particular sequence of flips, x_o, and based on these observations we wish to infer the probability of various values of θ. We are thus given two pieces of information

(i) There is some non-trivial relationship between x and θ which describe how a-priori plausible we consider the joint observation of two values of x and θ to be: $q(x, \theta)$.

(ii) After we have observed the data we *know* the data vector x takes a definite value x_o.

The information we observe a particular value of x_o affect our beliefs of θ. If for instance we observe $x = x_o =$ ("heads","heads","tails") then we cannot afterwards believe $x =$ ("tails","heads","tails"). If we denote our beliefs after observing x_o by the distribution $p_o(x, \theta)$ we *must* rule out those beliefs where x take another value than x_o. In particular it must hold for $p_o(x) = \int d\theta \; p_o(x, \theta)$:

$$p_o(x) = \delta_{x - x_o} \tag{3.86}$$

and we denote by $\mathcal{C}$ the distributions $p(x, \theta)$ which satisfy this constraint:

$$\mathcal{C} = \left\{ p : \int d\theta \; p(x, \theta) = \delta_{x - x_o} \right\}. \tag{3.87}$$

While clearly smaller than the set of all distributions on $\mathcal{X} \times \Theta$, this set of distributions contain all distributions with a density of the form $\delta_{x - x_o} \mu(\theta)$ where μ is an arbitrary density of θ. We can now apply the above machinery and use eq. (3.85) to assign an order of preference on all posterior distributions $p(x, \theta)$ subject to $\mathcal{C}$. In particular the most preferred distribution p_o is given by

$$p_o = \arg\max_{p \in \mathcal{C}} S[p, q]. \tag{3.88}$$

The set eq. (3.87) might appear slightly discomforting, however consider any given value of x, $x' \in \mathcal{X}$. If x' happens to be equal to x_o then a distribution $p \in \mathcal{C}$ should have support on x', $p(x') > 0$. On the other hand if $x' \neq x_o$ then $p(x') = 0$. We can write this as follows:

$$\text{for all } x' \in \mathcal{X} \text{: } p(x') = \delta_{x' - x_o} \tag{3.89}$$

Now recall $p(x') = \int d\theta \; p(x', \theta)$. We can re-write the left-hand side of the above expression in the following manner:

$$\text{for all } x' \in \mathcal{X} \text{: } \int dx d\theta \delta_{x - x'} p(x, \theta) = \delta_{x' - x_o}. \tag{3.90}$$

Notice the above has the form of one linear equality constraint for each $x' \in \mathcal{X}$. Optimizing eq. (3.85) under eq. (3.90) (and introducing one lagrange multiplier $\lambda_{x'}$ for each $x' \in \mathcal{X}$) we obtain the variational problem

$$\delta \left\{ S[p,q] + \alpha \left(\int dx d\theta \; p(x,\theta) - 1 \right) + \int dx' \; \lambda_{x'} \left(\int dx d\theta \delta_{x-x'} p(x,\theta) - \delta_{x'-x_o} \right) \right\}. \tag{3.91}$$

Taking the functional derivative and maximizing we arrive at

$$p(x,\theta) = \frac{1}{Z} q(x,\theta) \exp(\lambda_x) = q(\theta|x) \frac{1}{Z} q(x) \exp(\lambda_x) \tag{3.92}$$

where Z is a normalization constant. The constants λ_x can be fixed by inserting eq. (3.92) into eq. (3.90) to obtain

$$\frac{q(x')}{Z} \exp(\lambda_{x'}) = \delta_{x'-x_o}. \tag{3.93}$$

Inserting eq. (3.93) into eq. (3.92) results in

$$p(x,\theta) = \delta_{x-x_o} q(\theta|x). \tag{3.94}$$

This expression is clearly marginally equivalent to $q(\theta|x_o)$, consistent with simply applying Bayes rule.

It is apparent the above framework reduce to the maximum entropy principle Jaynes [1957a,b] if the linear constraints are of the usual sort and the prior q is chosen uniformly. However the above principle allow mixing constraints with observed data; it is by allowing such mixing it can be said to generalize the maximum entropy principle and Bayes updating [Caticha and Giffin, 2006].

While the form eq. (3.94) might appear ugly, the above treatment give a more accurate description of learning: In a standard description of Bayesian learning it is common to distinguish verbally between $q(\theta)$ and $q(x|\theta)$ (the prior and the likelihood), however here it is treated as a single object $q(x,\theta)$ which capture *all* relevant information available before observing data. After observing the data we are left with eq. (3.94); this object provide a better representation of what we know after observing the data than $p(\theta|x_o)$, namely how plausible different values of the parameters are and also what data was actually observed.

This result is encouraging in another way too. Suppose we first observe some data x_1 and then x_2. Denoting all data by $x = (x_1, x_2)$, the inference we should like to draw about the parameter θ is

$$q(\theta|x_1,x_2) \propto q(x_1|x_2,\theta) q(x_2|\theta) q(\theta) \tag{3.95}$$

Consider sequential updating of our beliefs q by first observing x_1 (to $q_1(\theta|x_1)$) and then x_2 (to $q_2(\theta|x_2)$). Omitting some details the updating based on eq. (3.11) would give:

$$q_2(\theta|x_2) \propto q(x_2|\theta)q_1(\theta|x_1) \tag{3.96}$$

$$\propto q(x_2|\theta)q(x_1|\theta)q(\theta). \tag{3.97}$$

This result show sequential updating is equal to Bayes updating eq. (3.95) in the case where the data is marginally independent conditional on the parameters.

While conditionally independent models such as the above play a very central role in machine learning, it is worth keeping in mind the derivation is not simply applying Bayes rule and in general there are cases where treating the constraints in different order give different results. In this case they are said to be *noncommuting*, see Caticha [2008], Caticha and Giffin [2006] for further discussion and an example.

3.4 Application of the MEP to Bayesian Dropout

So far we have argued under certain assumptions, desiderata (C1)–(C5), there exist a single unique order of preference for updating beliefs, namely the (negative) Kullback-Leibner divergence. However there still remains the question which constraints are relevant. For instance in the example of the coin in section 3.3.5 it is apparent no other constraint except for the (canonical) data constraints *should* be imposed. On the other hand, if we consider a situation where other constraints are relevant (such as an energy constraint as in an ideal gas[Jaynes, 1957a,b]), the theory certainly allow us to impose additional data constraints, but it is in this context hard to imagine a situation where we actually had such data and would be interested in the resulting posterior.

We suggest this might be due to simply looking at the wrong place for constraints. Consider the canonical situation where expectancy constraints of the form eq. (3.79) are relevant, e.g. energy constraint such as for the Ising spin model or in an ideal gas. If we try to see this in a machine learning perspective the particles in the ideal gas, i.e. their position and momentum, is what we would expect to measure. Accordingly we are easily lead into thinking we should look for a situation where we have (some) partial observation of data (position and momentum) and in addition a constraint also expressed on the data. However in this case the data and constraints compete for degrees of freedom, i.e. if we observe all the data an expectancy constraint such as eq. (3.79) expressed on the data will be fully determined and will either be irrelevant or lead to a conflicting state of information.

Posed in this way it is natural to not look at the data but at the *parameters* of the model as the place where constraints might apply. This lead to two lines of thought: Either the parameters represent something physical, for instance gender and days of week as in the example with the girl born on a Tuesday in chapter 2. In this case constraints should represent some physical fact in themselves, i.e. representing an energy or magnetization constraint. This bring us back to the situation outlined in the previous paragraph only with the additional complication statistical models do not usually come with parameters with such a rigid physical role (this is leaving out the additional problem from where we would know the value of the expectation, i.e. the right hand side of eq. (3.79)).

On the other hand the parameters might have a more free-floating interpretation, i.e. they are simply the way the model happens to be parameterized out of analytical convenience. However in this case why is it reasonable ignore *some* possible posteriors by applying constraints?

Before fully accepting this negative lesson it is important to consider what the model represents. The model too is simply a restriction on the class of possible posteriors and on this view there is little reason to prefer one type of restriction (a particular functional form of the model) over another (some sort of constraint). While much work in machine learning has focused on constraints of the first form, models, in the next section we will consider a technique which fall under the second form. The section is based on the work found in *Bayesian Dropout* [Herlau et al., 2015].

3.4.1 Dropout

Consider a completely different setting, namely a classical feed-forward neural network. A neural network attempt to model data y based on input x by adapting weights θ. In a typical neural network application the input-output relationship is complicated (for instance the input could correspond to natural images and the output could correspond to the identification of certain faces in the images) and so in the absence of strong prior information the parametrization must support many possible mappings. This is typically done by using several intermediate layers with many neurons.

This lead to the problem different settings of parameters will be able to model the input-output relationship perfectly while giving very different (poor) predictions on the test data. This problem is sometimes called co-adaptation because different coordinates of the parameter vector co-adapt to each other to give predictions specific for the training and not the test data [Hinton et al., 2012].

Dropout, originally proposed by Hinton, Srivastava, Krizhevsky, Sutskever, and Salakhutdinov [2012] has been proposed as a way to reduce such co-adaptation. Dropout is best explained as an online algorithm: A neural network is typically trained by gradient descent, that is, in each iteration the gradient (at the current setting of the parameters) is computed and the parameters are changed in the direction of the gradient. With dropout, at each iteration the parameter vector is altered by "turning off" a subset (chosen stochastically) of the parameters, typically by fixing their value to zero. We write this operation as:

$$\tilde{\theta} \leftarrow I \circ \theta$$

where I is a new (random) vector and $\circ$ is the operation where the parameters is perturbed. The coordinates not set to zero is then updated based on the gradient computed at $\tilde{\theta}$.

This technique limits co-adaptation in that no parameter, even on the same training data, will have access to the exact same configuration of other parameters as these are stochastically dropped out. The technique has shown to lead to significant performance gains (see for instance Hinton et al. [2012], Krizhevsky et al. [2012], Dahl et al. [2013]). In the remained of this section we will give it a probabilistic interpretation using the MEP.

It is natural to consider the following generative model for each observation y_i given x_i:

$$\theta \sim q(\cdot) \tag{3.98a}$$

$$\text{for each } i\text{: } I_i \sim q(\cdot) \tag{3.98b}$$

$$y_i | I_i, \theta, x_i \sim q(\cdot | I_i \circ \theta, x_i) \tag{3.98c}$$

Where it is assumed omitting I will give some valid neural-network like model. While this formulation is seemingly what we want from dropout, the role played by I is very different. In the current formulation I is adapted to each observation; that is to say, we *adapt* both the weights *and* which weights are dropped out to the particular observed data y_i. While doing this within a Bayesian framework will likely reduce the co-adaptation (ie. θ is not only selected as a single value but as a distribution over different values), it is clearly a very different approach than dropout where we actively *prevent* co-adaptation. There are various ways to formulate this discrepancy, the most pointed is to say there is something *"wrong"* in us learning

$$q(I_i | x_i, y_i, \theta). \tag{3.99}$$

This immediately suggest a solution: To say we should not learn (or co-adapt) I_i based on the data (x_i, y_i) and weights θ is exactly to say our posterior knowledge

(after observing the data) should be the same as our prior knowledge insofar I is concerned. Specifically:

$$\textit{Bayesian Dropout}: \qquad p(I_i|x_i, y_i, \theta) = q(I_i). \tag{3.100}$$

Now the idea is simply to update p from q based on data and the above constraint. The resulting posterior become [Herlau et al., 2015]

$$p(\theta) = \frac{1}{Z} q(\theta) \exp\left(\sum_i \sum_{I_i} q(I_i) \log q(y_i|x_i, I_i \circ \theta) \right). \tag{3.101}$$

This naturally leave a few questions open. For instance the above form will in general constitute a double-stochastic sampling problem which is not trivially solved. In Herlau et al. [2015] we discuss three different techniques for inference and illustrate them on linear model and logistic regression. The bottom line is there seem to be some benefit for the above technique when the number of data dimensions is large compared to the number of samples; exactly the situation where we would expect dropout to work Hinton et al. [2012]. Taking the logarithm of the above target, throwing away the prior terms and optimizing recovers the objective:

$$O(\theta) = \sum_i \sum_{I_i} q(I_i) \log q(y_i|x_i, I_i \circ \theta) \tag{3.102}$$

This objective function is implied by dropout from standard arguments form stochastic optimization in the small stepsize limit [Amari, 1997, Robbins and Monro, 1951] and has previously been independently proposed by amongst other Wang and Manning [2013], we will however not discuss the result further here and only mention it briefly in the conclusion in chapter 7.

Chapter 4

Symmetries and invariance

Consider the following situation: In 2024, scientists are trying to determine if a particular elementary particle (particle $\mathcal{P}$) exists. They attempt to do so in an experiment in which particles are smashed together in an accelerator and based on these independent collisions they attempt to draw inference on the particles existence. We imagine data analysis has changed by 2024, such that instead of considering p-values, the scientists have constructed a robot which attempts to compute $p(H|\cdots)$ where

$$H : \textit{"The particle } \mathcal{P} \textit{ exists."} \tag{4.1}$$

and the dots stand for the available evidence such as the result of collisions, physical theories and other relevant information. The robot is built to act in full accordance with the theory encountered in chapter 2. After one year of smashing particles the day has finally come where they ask the robot:

Scientist: Here is the data from the last year, Y_a, does particle $\mathcal{P}$ exist?

Robot: Particle $\mathcal{P}$ exists, $p(H|Y_a, \cdots) = 0.63$.

Scientist: Great! That should ensure funding for next years operations.

Robot: Particle $\mathcal{P}$ does not exist, $p(H|Y_a, \cdots) = 0.47$.

Scientist: That is absurd. We don't know anything about the potential collisions next year! I might as well have said we could have smashed particles for a million years starting tomorrow–

Robot: Particle $\mathcal{P}$ exists, $p(H|Y_a, \cdots) = 0.52$.

This behavior is certainly paradoxical, but it is not inconsistent with anything in chapter 2. To understand what happends we need to be more explicit: Let Y_a correspond to the data for the first year and Y_b the data for the next year. Ignoring any other information the key observation is the joint distributions are defined over different sets of variables: $p_a(H, Y_a)$, $p_{ab}(H, Y_a, Y_b)$ and so the marginal distributions:

$$p_a(H|Y_a) = \frac{p_a(H, Y_a)}{p_a(Y_a)} \tag{4.2a}$$

$$p_{ab}(H|Y_a) = \frac{\int dY_b \; p_{ab}(H, Y_a, Y_b)}{p_{ab}(Y_a)} \tag{4.2b}$$

need not be equal. Thus at the beginning of the dialogue the robots available data corresponded to the computation in eq. (4.2a), however when it learned there would be another year of (unobserved) data it rightly changed to the computation in eq. (4.2b), and similar when it considered a potentially infinite stream of data.

While what the robot does is formally possible, it is clearly not drawing the right inferences: We want the robot to *not care* about unobserved data. One way to state this requirement is to say the robot should act consistently in the sense:

$$p_a(H, Y_a) = \int dY_b \; p_{ab}(H, Y_a, Y_b). \tag{4.3}$$

An additional mistake the robot could possible make is if the order the experiments matter. Consider a new dataset Y_a' obtained by permuting the entries in Y_a, for instance by flipping the observation made last Tuesday with those made last Monday. Quite clearly the labelling information should carry no information (we assume the robot is considering fixed natural laws and equipment) and we expect the two distributions to agree as well. The notion of *exchangeability*, which we will discuss below, captures invariance under permutations and consistency in the sense of eq. (4.3).

More generally, this chapter will consider random *objects* (or *structures*) such as lists, arrays, partitions or trees as well as various probabilistic symmetries that are thought appropriate for each type of object. An example is a list (for

instance a list of observations as in the previous example) and the symmetry is exchangeability, that is, that the distribution of the object is invariant under permutations.

After identifying appropriate symmetries and objects, the most apparent problem is to characterize the class of objects for a given type which obey a particular symmetry, for instance all partitions that are exchangeable. In this section we will give some examples of structures and symmetries that will be used in the later work.

A brief note on formality It is worth emphasizing words like probability will now have a quite different meaning than in chapters 2 and 3. In the previous chapters the probability distribution (or simply the probability) was considered a function and in chapter 3 even an analytical function, however from a technical perspective probability theory is a subfield of measure theory.

Changing the underlying mathematical object between chapters is regrettable from a formal perspective, but it was necessary to discuss the different concepts. One way to overcome this tensions is to consider "probability theory" as what we arrived at in chapter 2 and measure theory as providing a particular mathematical model of probability theory realized in the formalism of measurable sets [Ballentine, 2001], however such an attempt require a long discussion on the meaning of "theory", "model" and is open to the charge the derivation in chapter 2 quite obviously assume a particular mathematical underpinning albeit not a very rigorous one. We will not discuss this issue further and simply assume probabilities are now understood in the context of measure theory.

We emphasize the following sections are not intended as a comprehensive guide to non-parametric methods but a brief review of some important results from the field. Good introductions to the field may be found in Kingman [1993] (point processes and random measures), Aldous [2010] (exchangeable random arrays, trees and other structures), Pitman and Picard [2006] (covering random partitions, Brownian motion, and coalescent with a focus on combinatorics), Kallenberg [2005] (likely the most comprehensive text on the subject but not an easy read). In addition to these texts Orbanz and Roy [2013] provide a recent review of non-parametric methods for machine learning with focus on the Aldous-Hoover theorem.

4.1 Exchangeable sequences

To properly express the results in this chapter we will need to distinguish between random variables and their value. To this end, let $(x_i) = (x_1, x_2, \dots)$ be a sequence of elements of a space $\mathcal{X}$, (x_{ij}) a two-dimensional array of elements of $\mathcal{X}$ and so on for higher-dimensional arrays (more properly, (x_{ij}) is a mapping $\mathbb{N}^2 \to \mathcal{X}$). When presenting the general results we will properly distinguish between a particular *value* x_i and the corresponding random variable X_i by using upper case letters.

Let (X_i) be a sequence of random variables over the same space $\mathcal{X}$. We call the sequence *exchangeable* if it satisfy

$$(X_1, X_2, \dots) \overset{\mathrm{d}}{=} (X_{\sigma(1)}, X_{\sigma(2)}, \dots) \tag{4.4}$$

for all finite permutations σ [Kallenberg, 2002, p. 168], [Aldous, 1985]. Before we relate the above expression to the previous discussion of the robot, we will make a few remarks on the notation. A finite permutation is a bijection $\sigma : \mathbb{N} \mapsto \mathbb{N}$ that only permute a finite number of elements, that is, for any permutation σ there is an integer $K_\sigma \geq 1$ such that $\sigma(i) = i$ for $i \geq K_\sigma$. All permutations will be finite in the following and the word will be omitted.

Secondly, the symbol $\overset{\mathrm{d}}{=}$ means *equal in law*. To convert the definition into more familiar territory, assume (A_i) is a sequence of measurable subsets of $\mathcal{X}$, for instance small intervals centered around a sequence of points $(a_i), a_i \in \mathcal{X}$. The statement: $X_i \in A_i$ is then the Boolean statement if the outcome of experiment i fall into interval A_i and the definition eq. (4.4) is [Orbanz and Roy, 2013]

$$P(X_1 \in A_1, X_2 \in A_2, \dots) = P(X_{\sigma(1)} \in A_1, X_{\sigma(2)} \in A_2, \dots). \tag{4.5}$$

We will follow the standard practice in using upper-case letters to denote the probability when it is understood in a measure-theoretical [Kallenberg, 2002].

For a discrete space $\mathcal{X}$ and in more standard (sloppy) notation eq. (4.5) would be

$$p(x_1, x_2, \dots) = p(x_{\sigma(1)}, x_{\sigma(2)}, \dots). \tag{4.6}$$

To connect this definition with the motivation, notice the outcome of a particular experiment $(x_i)_{i=1}^n$ is obtained by marginalizing out the (infinite) sequence of observations $(x_i)_{i=n+1}^\infty$ and it follows the distribution of the first n observations is consistent with that of the first $n+1$ observations with observation x_{n+1} marginalized out. Furthermore, the permutation ensures the distribution of any subsequence of (x_i) of length n is equal in law to the distribution of the first n observations.

A main result, originally due to De Finetti [1931] (see also de Finetti [1974]) who first showed the result for binary observations. The general statement given below was first proven in Hewitt and Savage [1955], see also Ressel [1985], Aldous [2010] and Orbanz and Roy [2013] for re-statements in more familiar notation.

Theorem 4.1.1 (de Finetti, Hewitt-Savage)**.** *Let (X_i) be a sequence of random variables in $\mathcal{X}$. The sequence (X_i) is exchangeable if and only if there exist a probability measure μ on the set of probability measures $M(\mathcal{X})$ on $\mathcal{X}$ such that for all n and all measurable $A_1, \ldots, A_n$:*

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{M(\mathcal{X})} \prod_{i=1}^{n} \theta(x_i)\mu(d\theta). \tag{4.7}$$

In addition, μ is the distribution function of the empirical measure, ie. if we define the empirical measure on $\mathcal{X}$ as

$$S_n(\cdot) \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot) \tag{4.8}$$

Then S_n converge to θ with probability 1:

$$S_n(A) \to \theta(A) \tag{4.9}$$

for all measurable A.

To put the first part of the theorem in familiar language we will first ignore the measure theoretic difficulties by for instance assuming $\mathcal{X}$ is discrete. The de Finetti theorem can then be put in the naive form: *There exist some high (possibly infinite)-dimensional object θ and probability distribution $p(\theta)$ such that each x_i is i.i.d. conditional on G with marginal distribution p_θ:*

$$p(x_1, \ldots, x_n) = \int d\theta \ \prod_{i=1}^{n} p_\theta(x_i)p(\theta). \tag{4.10}$$

The above form, where θ is considered a (potentially infinite) set of parameters for a parametric density p_θ (the particular parameterization is of course determined by the type of sequence) is often more intuitive and the text will often resort to this form.

4.1.1 Example: The normal mixture-model

Suppose we wish to model continuous observations in $\mathbb{R}^d$, for instance we can consider the outcome of a collision in the particle accelerator as being composed of d scalar measurements from d detectors, x_i. For fixed K the following

construction could be proposed as a starting point:

$$p(x_i|\theta) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(x_i|\mu_k, \Sigma_k) \tag{4.11a}$$

$$\mu_k \sim \mathcal{N}(0, I_d) \tag{4.11b}$$

$$\Sigma_k = \sigma_k^2 I_d \tag{4.11c}$$

$$\sigma_k \sim \text{Gamma}(1, 1). \tag{4.11d}$$

In this case notice $\theta \equiv \left((\mu_k)_{k=1}^K, (\sigma_k)_{k=1}^K\right)$ and according to de Finetti's theorem the above construction is exchangeable. It should be apparent it is easy to construct a *generative* process for an exchangeable process, however the above model is limited to a particular choice of K. While we might soften this requirement by letting K come from some distribution, our initial desire to model a *source* of observations (x_i) of arbitrary length gives rise to a dilemma: If we let this distribution select K from a simple distribution such as from a Poisson distribution, K may appear very conservative if the particular observed subsequence of (x_i) happens to be very long. On the other hand, if K is selected to have uniform support, $p(K) \propto 1$, what then is the generative process?

Solving these problems in the most general way that leads to tractable inference has been the focus of much research in probability theory and Bayesian nonparametrics in machine learning. If we accept the above intuition as having general applicability, it is easy to see the more general models are those obtained by letting K be large; it turns out again and again in different settings the principal way to select a parameter such as K is to let K be countably infinity and choose a parameterization which solves the problems which naturally arises. In the following sections we will briefly review some of these results.

4.1.2 Convergence

Consider again the de Finetti theorem 4.1.1. According to eq. (4.7) the data is explained by first drawing a random measure θ from a distribution μ over all measures $M(\mathcal{X})$. Thus, if we observe a particular sequence of data $(X_i)_{i=1}^n$ we can compute the posterior distribution of θ by conditioning on the data; to avoid measure-theoretical problems we will give the result in the naive notation eq. (4.10)

$$p(\theta|(x_i)_{i=1}^n) \propto \prod_{i=1}^{n} p(x_i|\theta)p(\theta). \tag{4.12}$$

The second part of de Finetti's theorem, eq. (4.8) implies that *if* the sequence $(x_i)_{i=1}^n$ *was actually generated* using eq. (4.7) from a particular random measure

θ drawn from μ then the posterior distribution will converge to a particular measure θ which is also the empirical measure. Notice this result assumes the data was generated from a particular distribution of random measure μ and it was this distribution we used as the prior when we applied Bayes theorem. Normally this cannot be expected to be the case nor would we *know* it was the case for actual data, and even if the posterior is concentrating the de Finetti theorem does not tell us how fast it is converging and in particular if it has converged by any meaningful standard on the particular data set.

In general, one should not assume the lesson from ordinary Bayesian inference, that is, any usual model with a prior of wide support will be sufficient to guarantee convergence if we are given enough data. The default position implied by the de Finetti theorem should be we do *not* have the correct model and thus we are *not* guaranteed convergence and convergence results have to be obtained by more elaborate arguments. In practice, non-parametric models appear well-behaved and convergence do not appear to be an issue. The interested reader is invited to consult Ghosal [2010] for a general introduction as well as Kleijn and van der Vaart [2006] for convergence in the misspecified case for (parametric) mixtures and nonparametric regression as well as Ghosal and van der Vaart [2007] for special cases involving Dirichlet mixture of normals and van der Vaart and van Zanten [2008], Castillo [2012] for special cases involving gaussian processes.

4.2 Exchangeable Partitions

To progress beyond the de Finetti theorem requires further assumptions on the data source. A particular important example is if the data represent a partition, in this case de Finetti's theorem take a particular simple form and the continuous representation θ in eq. (4.7) can be identified as what is called a *paintbox*. The notation will be important when discussing models for random hierarchies (trees) later in section section 4.4. The definitions in the following section are taken from [Kingman, 1978, Pitman and Picard, 2006]

Elementary definitions Recall a partition π of a set $\mathcal{X}$ is a collection of subsets $B_1, \ldots, B_K$ of $\mathcal{X}$ such that for all $1 \leq \ell < m \leq K$ (K may be infinite):

$$B_\ell \neq \emptyset, \qquad \bigcup_{\ell=1}^{K} B_\ell = \mathcal{X}, \qquad B_\ell \cap B_m = \emptyset \tag{4.13}$$

written as $\pi = \{B_1, \ldots, B_K\}$ and each B_ℓ is called a *block* and we use the notation $b \in \pi$ to indicate b is a block in π, ie. $b = B_\ell$ for some $1 \leq \ell \leq K$.

We will write $\Pi_{\mathcal{X}}$ for a random partition of the set $\mathcal{X}$. In the following we will be particularly interested in the case where $\mathcal{X}$ is discrete. Since the labelling of the elements in $\mathcal{X}$ is assumed unimportant, we use the shorthand $\Pi_n \equiv \Pi_{[n]}$. For a partition $\pi = \{B_1, B_2, \ldots, B_K\}$ of $\mathcal{X}$ and a permutation σ of $\mathcal{X}$ we may define the action of σ on π by

$$\sigma(\pi) \equiv \{\sigma(B) : B \in \pi\} \tag{4.14}$$

$$\sigma(A) \equiv \{\sigma(i) : i \in A\}. \tag{4.15}$$

A random partition Π_B is then called exchangeable if for all permutations σ:

$$P(\Pi_B = \pi) = P(\Pi_B = \sigma(\pi)). \tag{4.16}$$

This is in turn equivalent to saying the distribution of $\Pi_{\mathcal{X}}$ only depend on the size (and not order) of the blocks of the partition, i.e. there is a symmetric function $p_{|\mathcal{X}|}$ such that

$$P(\Pi_{\mathcal{X}} = \pi) = p_{|\mathcal{X}|}(|B_1|, \ldots, |B_K|) \tag{4.17}$$

where p is symmetric in its arguments. This function is called the *exchangeable partition probability function* (EPPF) [Pitman, 1995].

Assume $A, B \subset \mathcal{X}$ (in the following it will always be the case that $A, B \subset \mathbb{N}$). Suppose π is a partition of B and $A \cap B \neq \emptyset$. We define the *projection* of the partition z onto $A \cap B$ by

$$\text{proj}_A(\pi) \equiv \{b \cap A : b \in \pi,\ b \cap A \neq \emptyset\} \tag{4.18}$$

Clearly the projection is also a partition of $B \cap A$.

Suppose $(\Pi_n) = (\Pi_1, \Pi_2, \Pi_3, \ldots)$ is an infinite sequence of exchangeable random partitions on $([n]) = ([1], [2], [3], \ldots)$. The sequence is called *projective* if for all m:

$$P(\text{proj}_{[m]} \Pi_n = \pi) = P(\Pi_m = \pi). \tag{4.19}$$

That is, the behavior of a partition of m elements is the same as the behavior of a partition of n elements restricted to a subset of m elements.

A fundamental result due to Kingman [1978] allow us to get rid of the infinite sequence of random partitions (Π_n) in the definition above without loss of generality.

Theorem 4.2.1 (Kingman)**.** *An infinite sequence* (Π_n) *of finitely exchangeable random partitions is projective iff. there is an exchangeable random partition of* $\mathbb{N}$*,* Π_∞*, such that for any* $B \subset \mathbb{N}$*:* $\Pi_B \overset{\mathrm{d}}{=} proj_B \Pi_\infty$*.*

It follows without loss of generality we can exclusively consider exchangeable random partitions Π_∞ of $\mathbb{N}$. With these definitions in place we are ready to consider the equivalent of de Finetti's theorem (theorem 4.1.1) for partition-type data. This requires an identification of the partition Π_∞ with the the sequence (X_i) on the left-hand side of eq. (4.7) and determine a parameterization of the random measure θ.

Kingman [1978] gave an example of such a parameterization known as a paint-box: First, the *space* of parameters consist of all infinite ordered sequences $\theta = (t_1, t_2, \dots)$ such that each $t_i \in [0, 1]$ and satisfy:

$$\sum_i t_i \leq 1 \qquad t_1 \geq t_2 \geq t_3 \geq \dots . \tag{4.20}$$

The reader should have the following picture in mind: Consider a stick of unit length. The stick is divided (starting from the left and moving to the right) into one interval covering $I_1 = [0, t_1[$, then another interval covering $I_2 = [t_1, t_1 + t_2[$ a third covering $I_3 = [t_1 + t_2, t_1 + t_2 + t_3[$ and so on. Since the length in general may sum to less than one there might remain some last part of the stick I_∞. In general we define:

$$T_k = \sum_{i=1}^{k} t_i, \qquad I_k = [T_{k-1}, T_k[, \qquad I_\infty = (1 - T_\infty, 1]. \tag{4.21}$$

The sets $(I_k)_k$ and I_∞ should be thought of as the boxes in the paint-box. Each interval of the stick (a box) is assigned a particular color chosen at random; let the color for interval I_k be c_k chosen at random from some space; to give an explicit example, consider the case where the colors are chosen as RGB coordinates:

$$\text{for } k \in \mathbb{N}\text{: } c_k \sim \text{Uniform}([0, 1]^3). \tag{4.22}$$

The last box, I_∞, is for now assumed to be uncolored. The mixture distribution p_θ is then defined as drawing a random number $U_i \sim \text{Uniform}([0, 1])$ and setting

$$X_i = \begin{cases} c_k & \text{if } U_i \in I_k \\ c & \text{if } U_i \in I_\infty \text{ and } c \sim \text{Uniform}([0, 1]^3). \end{cases} \tag{4.23}$$

Clearly by the de Finetti theorem this induce an exchangeable random *coloring* (X_i), however each realization (x_i) of (X_i) induces an infinite partition π by

the convention i, j, $i \neq j$ are in the same block in π iff. $x_i = x_j$. As before we denote by Π_∞ the induced random partition corresponding to (X_i) and we will call both (X_i) and Z_∞ the the paint-box distribution with parameter θ.

The fundamental result is now summarized in the following theorem: [Kingman, 1978]

Theorem 4.2.2 (Kingman)**.** *Let Π_∞ be a random partition. Π_∞ is exchangeable if and only if it has the same distribution as the partition structure induced by the paint-box construction*

$$P(\Pi_\infty \in \cdot) = \int d\theta \prod_{i=1}^{\infty} p_\theta(\cdot)\mu(d\theta) \tag{4.24}$$

for some μ on the set of infinite sequences θ and p_θ being the paint-box distribution. Furthermore, if Π_∞ is exchangeable and for each fixed n the sequence $(N^{\downarrow}_{n,i})$ is the decreasing rearrangement of block sizes of Π_n with the convention $N^{\downarrow}_{n,i} = 0$ if Π_n has fewer than i blocks then the block sizes t_i in eq. (4.20) *may almost surely be recovered as*

$$t_i = \lim_{i\to\infty} \frac{N^{\downarrow}_{n,i}}{n} \tag{4.25}$$

Notice the two parts of theorem 4.2.2 correspond to the two parts of de Finetti's theorem eqs. (4.7) and (4.8). Secondly, notice a draw X_i from the paint-box construction parameterized by θ can simply be written as:

$$X_i \sim p_\theta(\cdot) \equiv (1 - T_\infty)\text{Uniform}([0,1]^3) + \sum_{k=1}^{\infty} s_k \delta_{c_k}. \tag{4.26}$$

It is the right-hand side of the above expression which can be substituted for the random measure in eq. (4.7). Consistent with chapter 2 we will use the shorthand $z_i \equiv \ell$ for a partition $\pi = \{B_1, \ldots, B_K\}$ to indicate $i \in B_\ell$, i.e. the index of the block i belong to.

What we have not specified is the distribution over θ. Once again we find ourselves in the situation the theory suggest nearly everything is possible as far as arriving at a proper generative model, however if we (as we would nearly always do) ask for the probability of any particular partition $P(\Pi_n = \{B_1, B_2, \ldots, B_k\})$ most suggestions would lead to analytical inconveniences. To proceed further we need to study particular choices of θ in de Finetti's theorem eq. (4.7) and this will be the goal of the next section.

4.2.1 The Dirichlet Process

Random measures play a crucial role for exchangeable structures, both in the de Finetti theorem (theorem 4.1.1) and in Kingmans paint-box construction (theorem 4.2.2). In this section we will introduce a particular construction for random measures, the *Dirichlet process*, which, while arguably not the simplest construction or most fundamental construction, is certainly the construction that has played the largest role in Bayesian non-parametrics for machine learning.

Consider again the setting of the de Finetti theorem (theorem 4.1.1) where we consider an exchangeable sequence of random variables (X_i), each taking values in a space $\mathcal{X}$, and the implied representation make use of a random measure p_θ (or simply θ) on the right-hand side of eq. (4.7). For the Dirichlet process this measure is normalized, ie. $p_\theta(\mathcal{X}) = 1$. The reader is encouraged to consider $\mathcal{X}$ to be a connected subset of $\mathbb{R}^2$ for convenience.

A normalized random measure can informally be thought of as follows: Suppose we have a small machine with a button. When we press the button, the machine produce a measure, $p_\theta(\cdot)$, of $\mathcal{X}$. This measure behaves just as any other measures in that for each measurable subset $A \subset \mathcal{X}$ we can compute the number $p_\theta(A) \in [0, 1]$. The machine is the equivalent of a random measure, that is, an object which (randomly) produces measures.

Consider the measure $p_\theta(\cdot)$ again and suppose we wish to know what p_θ does but we do not have access to its analytical form. What we can do is to evaluate p_θ on various sets. Suppose we partition $\mathcal{X}$ into a large number K of small pieces $A_1, \ldots, A_K$. In the following discussion the reader should think of these pieces as fixed but arbitrarily chosen, and in the example of the subset of $\mathbb{R}^2$ the reader can imagine a fine grid. If K is very large, the vector

$$w \equiv \big(p_\theta(A_1), p_\theta(A_2), \ldots, p_\theta(A_K)\big) \in [0, 1]^K \tag{4.27}$$

will now tell us *nearly* all there is to know about p_θ. Notice $\sum_{k=1}^K w_k = 1$ since the measure is normalized. Suppose we keep the particular partition $(A_k)_{k=1}^K$ fixed and press the button N times on the machine to produce N measures: $p_\theta^{(i)}$, $i = 1, \ldots, N$. Evaluating eq. (4.27) now give N K-dimensional vectors $w^{(i)}$; in the same sense eq. (4.27) tell us *nearly* all there is to know about a single measure p_θ when K is large, the set of N vectors $(w^{(i)})_{i=1}^N$ tells us *nearly* all there is to know about the random measure (or *machine*) when K and N are both large.

To proceed it is natural to consider the mean and variance of $(w^{(i)})_{i=1}^N$:

$$\overline{w}_N = \frac{1}{N}\sum_{i=1}^{N} w^{(i)} \qquad \mathrm{Var}[w] = \frac{1}{N}\sum_{i=1}^{N}(w^{(i)})^2 - \overline{w}^2. \tag{4.28}$$

where both are understood as K dimensional vectors. Now, consider the case where the above machine, μ, enters into the de Finitti representation eq. (4.27). Then for any set A and a sequence of (just) one variable X_1:

$$p(X_1 \in A) = \int d\theta \; p_\theta(A)\mu(d\theta) \tag{4.29}$$

By exchangeability, we are justified in calling the right-hand side of eq. (4.29) the *mean* distribution of this particular random measure evaluated on A. It is in other words a characteristic of the machine that produced measures. Denote this mean by H To be completely explicit, H is the normalized measure on $\mathcal{X}$ defined as

$$H(A) \equiv \int d\theta \; p_\theta(A)\mu(d\theta) \tag{4.30}$$

for any $A \subset \mathcal{X}$. Clearly we can evaluate H on the partition $(A_k)_{k=1}^K$; it is not hard to see

$$\lim_{N\to\infty} \overline{w}_N = (H(A_1), H(A_2), \ldots, H(A_K))$$

Notice we have so far not put any limitations on the machine which produced random measure, μ: H is simply a property any such well-behaved machine must have and for any such machine we can compute the set of vectors $w^{(i)}$ which will have some distribution. Now, suppose we ask what distribution this particular set of vectors have. If we are optimistic we could hope this distribution was simple; the simplest case we could hope for was the Dirichlet distribution. That is, there is an $\alpha > 0$ such that

$$(w_1, w_2, \ldots, w_K) \sim \mathrm{Dirichlet}\left(\alpha H(A_1), \alpha H(A_2), \ldots, \alpha H(A_K)\right). \tag{4.31}$$

Notice the average of this Dirichlet distribution come out correctly as $H(A_1), \ldots, H(A_K)$ regardless of α. The limits $\alpha \to 0$ and $\alpha \to \infty$ should be kept in mind and will be consistent with the basic properties of the Dirichlet distribution. The former correspond to maximal variance and the later to a variance of 0.

The preceding discussion has assumed (A_k) was fixed and eq. (4.31) simply represented a particular possibility for how a particular machine μ behave on this particular set (A_k). Any fixed partition is only sufficient to give partial characterization of the underlying random measure, however the following definition due to Ferguson [1973] overcomes this limitation

Definition 4.2.1 (Dirichlet Process)**.** *Assume H is a fixed distribution over $\mathcal{X}$ and $\alpha > 0$ a real number. We say a random measure $\theta(\cdot)$ on $\mathcal{X}$ is distributed as a Dirichlet Process (DP) with base measure H and concentration parameter α if for any finite measurable partition $A_1, \ldots, A_K$ of $\mathcal{X}$*

$$(\theta(A_1), \ldots, \theta(A_K)) \sim \text{Dirichlet}\left(\alpha H(A_1), \ldots, \alpha H(A_K)\right). \tag{4.32}$$

This is written as $\theta \sim DP(\alpha, H)$.

If $\mathcal{X}$ is non-trivial, the number of measurable partition will be uncountable and so eq. (4.32) will consist of an uncountable number of constraints on the DP. Thus it is natural to ask at least three questions: Firstly, if such an object can exist, secondly, if it can be represented in a sensible and useful way and thirdly, if it allows simple computation of posterior predictive distributions and other useful quantities. Fortunately, the answer to all these question are yes. Firstly, there are several ways to establish existence first discussed by Ferguson [1973] under certain regularity conditions on H and $\mathcal{X}$, see also Blackwell and MacQueen [1973]. These restrictions can be softened significantly by the construction given by Sethuraman [1991].

4.2.1.1 Posterior Distribution

Suppose $(x_i)_{i=1}^n$ is a sample drawn from a Dirichlet process. I.e. we first draw $\theta \sim DP(\alpha, H)$ and then draw X_i i.i.d. from θ. Assume again $(A_k)_{k=1}^K$ is a finite partition of $\mathcal{X}$. Consider the posterior distribution implied by a standard application of Bayes theorem and the definition in eq. (4.32):

$$\begin{aligned} &p(\theta(A_1), \ldots, \theta(A_K) | x_1, \ldots, x_n) \\ &\propto \left[\prod_{i=1}^{n}\prod_{k=1}^{K} \theta(A_k)^{1_{A_k}(x_i)}\right] \text{Dirichlet}(\theta(A_1), \ldots, \theta(A_K) | \alpha H(A_1), \ldots, \alpha H(A_K)) \end{aligned} \tag{4.33}$$

which implies

$$\begin{aligned} &p(\theta(A_1), \ldots, \theta(A_k) | x_1, \ldots, x_n) \\ &= \text{Dirichlet}(\theta(A_1), \ldots, \theta(A_K) | \alpha H(A_1) + n_1, \ldots, \alpha H(A_K) + n_K). \end{aligned} \tag{4.34}$$

Where $n_k = |\{i : x_i \in A_k\}|$ and $1_{A_k}(x_i)$ is equal to 1 if $x_i \in A_k$ and 0 otherwise. Since the above holds for all partitions (A_k) of $\mathcal{X}$ the posterior distribution must be a DP too. Notice n_k depend on A_k. Rearranging eq. (4.34) shows

that the posterior DP can equally well be written as a new DP with updated concentration parameter and base measure:

$$(\alpha, H)|x_1, \ldots, x_n = \left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{x_i}}{\alpha + n}\right). \tag{4.35}$$

The Dirichlet process is thus conjugate under posterior updates, and this should naturally make us suspect the predictive posterior distribution, that is, the density of X_{n+1} conditional observations $x_1, \ldots, x_n$, should be simple. Indeed, the posterior of $\theta(A), \theta(\mathcal{X} \setminus A)|x_1, \ldots, x_n$ is given by eq. (4.35) and thus the posterior can be found as simply:

$$\begin{aligned} P(X_{n+1} \in A|x_1, \ldots, x_n) &= \int d\theta\ \theta(A) P(\theta|x_1, \ldots, x_n) \\ &= \frac{1}{n+\alpha}\left(\alpha H(A) + \sum_{i=1}^{n} \delta_{x_i}(A)\right). \end{aligned} \tag{4.36}$$

In other words, assuming $\mathcal{X}$ is large, with probability one it will be true with probability $n/(n+\alpha)$ the new observation x_i will be equal to one of the previous $x_1, \ldots, x_n$. Assume there are K unique values of x_i and denote these by $(x_k^*)_{k=1}^K$. Letting $n_k = |\{j : x_j = x_k^*\}|$ then eq. (4.36) simply becomes:

$$x_{n+1}|x_1, \ldots, x_n \sim \frac{\alpha}{n+\alpha} H + \sum_{k=1}^{K} \frac{n_k}{n+\alpha} \delta_{x_k^*}. \tag{4.37}$$

Thus we can easily obtain a sample $x_1, \ldots, x_n$ from a $DP(\alpha, H)$ process by using eq. (4.37) n times and, having obtained such a sample the likelihood will be:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) \tag{4.38}$$

where each probability is simply obtained from eq. (4.37). The reader should assure himself both of these procedures are easy to implement on a computer. Now, with a bit of tedious algebra it should be easy to see that, taking eq. (4.37) together with eq. (4.38) as the *definition* of a particular distribution of n data points and rearranging one obtains:

$$\begin{aligned} p(x_1, \ldots, x_n) &= \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) \\ &= \prod_{i=1}^{n} p(x_{\sigma(i)}|x_{\sigma(1)}, \ldots, x_{\sigma(i-1)}) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \end{aligned} \tag{4.39}$$

for any permutation σ. Since this is true for any n the sequences generated by eqs. (4.37) and (4.38) will be exchangeable. We can then invoke de Finetti's theorem to conclude there exists a random variable $\tilde{\theta}$ such that any sequence (X_i) is composed of iid. draws

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int \mu(d\tilde{\theta}) \prod_{i=1}^{n} \tilde{\theta}(A_k). \tag{4.40}$$

Assume a very long sequence $(x_i)_{i=1}^n$ was generated from the de Finetti representation (4.40) using a particular $\tilde{\theta}$. Since, per assumption, $\tilde{\theta}$ was drawn from a $DP(\alpha, H)$ process eqs. (4.35) and (4.37) can also be used to describe the posterior measure thus we get the measure μ in eq. (4.40) must be the Dirichlet process establishing existence. A rigorous version of this argument was first used by Blackwell and MacQueen [1973] to show the existence of the Dirichlet process.

4.2.1.2 The Dirichlet process and clustering

As already noted, the posterior representation of the Dirichlet process eq. (4.37) for n observations has the property of grouping objects together into K groups each of n_k objects and this induces an exchangeable partition. Suppose we are only interested in the partition structure, as a direct consequence of exchangeability and eq. (4.38) the density of any given partition π becomes independent of H:

$$p(\pi|\alpha) = p(n_1, \ldots, n_K|\alpha) = \frac{\Gamma(\alpha)\alpha^K}{\Gamma(n+\alpha)} \prod_{k=1}^{K} \Gamma(n_k). \tag{4.41}$$

Any partition with this density can be generated in an analogous fashion to eq. (4.37), namely by (i) starting with a single element (ii) adding a new element to a new block with probability proportional to α or an existent block with probability proportional to n_k. This process is known as the *Chinese restaurant process* and its study, originally motivated in the study of population genetics, predate that of the Dirichlet process [Ewens, 1972, Aldous, 1985]. In the following we will write $\pi \sim \mathrm{CRP}(B, \alpha)$ to indicate π is a partition of a set B distributed as a Chinese restaurant process eq. (4.41) and if B is assumed known we will abbreviate this as $\pi \sim \mathrm{CRP}(\alpha)$.

Let us recap some of the properties of the Dirichlet process discussed so far. Starting from the abstract definition of the Dirichlet process eq. (4.31) defined in terms of its action on any partition we have shown the Dirichlet process must be (i) exchangeable and (ii) give rise to posterior updates of the form eq. (4.37) and (iii) the posterior updates naturally induces a clustering, the CRP.

What we have so far not seen is the actual representation of the underlying infinite-dimensional measure θ implied by the de Finetti theorem. This construction was given by Sethuraman [1991] and turn out to be very simple. Assume $\alpha > 0, 0 \leq d < 1$ and for $k = 1, \ldots, \infty$:

$$V_k \sim \text{Beta}(1 - d, \alpha + kd) \qquad x_k^* \sim H \tag{4.42a}$$

$$\pi_k = V_k \prod_{\ell=1}^{k-1} (1 - V_\ell) \qquad \theta = \sum_{k=1}^{\infty} \pi_k \delta_{x_k} \tag{4.42b}$$

Then if $d = 0$, the generated vector θ is distributed according to a $\text{DP}(\alpha, H)$ process (the case of general $0 < d < 1$ will be discussed below). Notice $\sum_k \pi_k = 1$ with probability 1, thus by sorting $(\pi_k)_k$ the above construction thus also become identical to the paint-box representation we know must exist by theorem 4.2.2. It follows the length of the sticks must also determine limiting frequency of the clusters in the CRP. As already noted, from a formal perspective the above construction extends the range of spaces H that allow a DP [Sethuraman, 1991].

4.2.2 Beyond the Dirichlet process

In the study of exchangeable observations, the Dirichlet process is undoubtedly the most widely applied object in modern Bayesian non-parametric modelling. Roughly speaking, the preceding sections has presented three approaches to constructing the Dirichlet process: (i) The characterization of the Dirichlet process by its action on subsets in eq. (4.32) (ii) The Chinese restaurant metaphor, ie. as a description of how one new observation x_{n+1} is generated from existing observations $x_1, \ldots, x_n$ as in eq. (4.37) (iii) As a prior over sticks using the construction in eq. (4.42).

Though the discussion has been somewhat interweaved in the preceding sections, each of these constructions could have been taken as a starting point and we would have ended up with roughly the same construction, the Dirichlet process. Notice the three descriptions offer different advantages and disadvantages, the most obvious is how to generate data. The construction (ii) is entirely trivial to generate samples from as it is contained in the definition. The construction (iii) is somewhat more difficult due to the inconvenience of having an infinite number of sticks, however if the sticks are generated dynamically it too would be fairly easy to implement. The approach (i) stands out in this regard as requiring one to first derive a representation before samples can be generated. On the other hand it should be apparent approach (iii), being nearly identical to a paint-box representation, offers easy generalizations while (i), being independent of the representation, might appeal the most to a mathematician.

These three approaches persist when we consider generalizations of the Dirichlet process. When considering generalizations from a machine-learning perspective we typically attempt to arriving at a more flexible model while still maintaining tractability, both analytical and computational. One of the most successful approaches which archive both aims is the two-parameter Poisson-Dirichlet process first proposed by Perman, Pitman, and Yor [1992] and further examined in various other papers Pitman and Yor [1997], Pitman [1995]. In the machine learning literature the process was popularized by Ishwaran and James [2001] as the Pitman-Yor process in the notation of a stick-breaking construction and may be obtained from eq. (4.42) by considering general $0 \leq d < 1$. The density of the induced partition, i.e. the direct generalization of eq. (4.41) for block-sizes $n_1, \ldots, n_K$ is the *two-parameter Chinese restaurant process* [Pitman, 1995]. In the case $\alpha \geq 0$ (notice this is not the most general range of α, c.f. [Pitman and Picard, 2006]):

$$p(z|\alpha, d) = p(n_1, \ldots, n_K|\alpha, d) = \frac{\alpha^K}{(\alpha - 1)^{(1)}} \left(\frac{\alpha}{d}\right)^{(n)} \prod_{b \in z} (-d)^{(|b|)} \tag{4.43}$$

$$\text{where:} \quad x^{(k)} = \frac{\Gamma(k + x)}{\Gamma(1 + x)} \tag{4.44}$$

A more general class of processes is known as the Gibbs-type priors which, to our knowledge, was first introduced by Gnedin and Pitman [2006] and can be seen as falling under the approach (ii) of proposing a different rule for generating a new observation. Aside the Pitman-Yor process, Gibbs-type priors contain the normalized inverse Gaussian process [Lijoi et al., 2005] and the normalized generalized gamma process [Lijoi et al., 2007b] as special cases [De Blasi et al., 2013]. These processes was also shown to intersect with the newly introduced extended Poisson-Gamma class of priors, a two-parameter family which admit a stick-breaking construction in the vein of (iii) which was recently introduced by James [2013].

The preceding list is by no means meant to be exhaustive and is by a large not very well explored from a computational perspective in the machine-learning literature. Some examples of notable applications which go beyond the Dirichlet process can be found within survival analysis[Jara et al., 2010],linguistics [Teh, 2006], topic modelling [Teh and Jordan, 2010] and biology[Lijoi et al., 2007a, Navarrete et al., 2008].

4.2.3 Completely random measures

Lastly we will consider an extension which we will loosely associated with approach (i), thought this requires some qualifications. Recall approach (i) at-

tempted to characterize the random probability measure by its distribution when evaluated on arbitrary measurable subsets of $\mathcal{X}$, see eq. (4.32). A generalization is to study a random *measure* which, as the name suggests, is a distribution over measures. The idea being if $\mu \sim \Theta$ is a random measure over over some space $\mathcal{X}$ then the measure $\tilde{\mu}$ defined by

$$\tilde{\mu} : A \mapsto \frac{\mu(A)}{\mu(\mathcal{X})} \tag{4.45}$$

for any set $\mathcal{X}$ will act as a random probability measure provided $0 < \mu(A), \mu(\mathcal{X}) < \infty$. To illustrate this construction, suppose x_1, x_2 and x_3 are three ordinary independent random numbers and consider the case where $x_1, x_2, x_3 \sim \text{Gamma}(A, 1)$ i.i.d. In this case the distribution of the normalized vector

$$\overline{x} = \frac{1}{x_1 + x_2 + x_3}[x_1, x_2, x_3]^T \tag{4.46}$$

follows a Dirichlet(A, A, A) distribution. By way of analogy, if we consider a random measure μ and a partitions A_1, A_2, A_3 of $\mathcal{X}$, then the random variables $X_1 \equiv \mu(A_1), X_2 \equiv \mu(A_2)$ and $X_3 \equiv \mu(A_3)$, are *independent* but can be made dependent through the normalization procedure eq. (4.45), and if the variables X_1, X_2, X_3 are Gamma$(A, 1)$ then the normalization procedure would behave as a Dirichlet process for this partition. Indeed, it has been argued most classes of random probability measures treated in modern Bayesian non-parametrics can be derived trough a suitable transformations of a particulary simple random measure known as a *completely random measure* [Lijoi and Prünster, 2010].

In the remainder of this section we will highlight a few important results for completely random measures beginning with the definition:

Definition 4.2.2 (Completely random measure)**.** *A* completely random measure *(CRM) on $\mathcal{X}$ is a random function μ from the collection of measurable subsets of $\mathcal{X}$ into $[0, \infty]$ such that (i) $\mu(\emptyset) = 0$ (ii) $\mu(A) < \infty$ for any bounded measurable subset $A \subset \mathcal{X}$ and (iii) for any countable partition $A_1, A_2, \ldots$ of $\mathcal{X}$ the random variables $(\mu(A_k))_k$ are independent and*

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k). \tag{4.47}$$

Or put more simply, a CRM is a random measure such that it gives rise to independent random variables $\mu(A_1), \ldots, \mu(A_k)$ when evaluated on disjoint subsets $A_1, \ldots, A_k \subset \mathcal{X}$.

As was the case for the Dirichlet process [Blackwell and MacQueen, 1973], a realization of a completely random measure is almost surely discrete and can be

represented as [Kingman, 1967]

$$\mu = \mu_0 + \sum_{i=1}^{\infty} w_i \delta_{\theta_i} + \sum_{i=1}^{\infty} v_i \delta_{\phi_i} \tag{4.48}$$

(compare to eq. (4.37)) where μ_0 is a fixed measure, $(w_i, \theta_i)_i$ is a *random* sequence in $\mathcal{X} \times \mathbb{R}^+$, (v_i) is a *random* sequence in $\mathbb{R}^+$ and $(\phi_i)_i$ is a *fixed* (non random) sequence in $\mathcal{X}$. In addition the sequences (v_i) is independent of the other quantities. The non-random measure μ_0 will be assumed to be zero in the following.

The sequence $(w_i)_i$ are commonly denoted the *random masses* and (θ_i) the *random locations* and each element (w_i, θ_i) an *atom*.

The second sum in eq. (4.48), where the locations are fixed, is important when characterizing the posterior of a CRM (compare to the fixed sequence in eq. (4.37)), but is often ignored when considering the CRM as a prior. Suppose then $v_i = 0$ for all i. A CRM μ is then characterized by the *Lévy-Khintchine* representation which states there exists a measure ν on $\mathbb{R}^+ \times \mathcal{X}$ obeying

$$\int_{\mathbb{R}^+} \int_{B \subset \mathcal{X}} \min[s, 1]\, \nu(ds, dx) < \infty \tag{4.49}$$

for all measurable, bounded B and such that for any measurable function $h : \mathcal{X} \mapsto \mathbb{R}$ it holds

$$\mathbb{E}\left[e^{-\int_{\mathcal{X}} \mu(dx) h(x)}\right] = \exp\left(-\int_{\mathbb{R}^+ \times \mathcal{X}} \nu(ds, dx)\left[1 - e^{-sh(x)}\right]\right) \tag{4.50}$$

assuming $\int |f| d\mu < \infty$ (with probability 1). The characteristic measure ν is commonly denoted the *Lévy-intensity* of the CRM μ or simply the *intensity*. For proof and further discussion see Kingman [1967, theorem 2].

Notice the important special case where $h(\cdot) \equiv u 1_A(\cdot)$ is the indicator function on a bounded set A scaled with $u > 0$. In this case eq. (4.50) reduces to

$$\mathbb{E}\left[e^{-u\mu(A)}\right] = \exp\left(-\int_{\mathbb{R}^+} \nu(ds, A)\left[1 - e^{-su}\right]\right) \tag{4.51}$$

that is, the Laplace transform of the random variable $\mu(A)$. Accordingly the representation allows amongst other things to compute the mean and other moments of $\mu(A)$ for all measurable $A \subset \mathcal{X}$ assuming these are finite.

To carry out the normalization procedure eq. (4.45) in a rigorous manner requires the denominator, $\mu(\mathcal{X})$, to be greater than zero and finite with probability

1. This can be ensured by the conditions

$$\int_{\mathbb{R}^+ \times \mathcal{X}} \nu(ds, dx) = \infty, \tag{4.52a}$$

$$\int_{\mathbb{R}^+ \times \mathcal{X}} \left[1 - e^{-us}\right] \nu(ds, dx) < \infty \text{ for all positive } u. \tag{4.52b}$$

then the corresponding CRM μ characterized through eq. (4.50) satisfies

$$0 < \mu(\mathcal{X}) < \infty \tag{4.53}$$

almost sure[Regazzini et al., 2003]. When the conditions eq. (4.52) are met for an intensity ν the normalization construction eq. (4.45) can be made rigorous, i.e. we define $\tilde{\mu} \equiv \frac{\mu}{T}$ where $T = \mu(\mathcal{X})$ and μ is a CRM with intensity ν and $\tilde{\mu}$ is denoted a *normalized random measure with independent increments* (NRMI) and will again be almost surely discrete [James, 2003], that is, the NRMI has a representation:

$$\tilde{\mu} = \sum_{i=1}^{\infty} \tilde{w}_i \delta_{\theta_i} \tag{4.54}$$

for sequences $(\tilde{w}_i)_i$ in $\mathbb{R}^+$ and $(\theta_i) \in \mathcal{X}$ (compare this to the stick-breaking representation of the CRP given in eq. (4.42)).

4.2.3.1 The Poisson process and completely random measures

The references in the preceding section argues many non-parametric random priors can be constructed through suitable transformation of a CRM [Lijoi and Prünster, 2010], for instance the normalization eq. (4.54) which allows the construction of random probability measures. It was also shown a CRM can be characterized through the intensity measure ν (eq. (4.50)), however this characterization itself does not allow us to carry out common operations such as sampling a CRM, i.e. obtaining the random masses and atoms $(w_i, \theta_i)_i$ in the representation eq. (4.48) (or at least an arbitrarily large subset of these). This limitation can however be overcome using another random process, the *Poisson process*, which can in turn be seen as a particulary instance of a CRM. A Poisson process (PP) Π on a set $\mathcal{X}$ is a random (countable) subset of $\mathcal{X}$. Suppose Π is a (countable) subset of $\mathcal{X}$. Then for any measurable $A \subseteq X$ define

$$N(A) \equiv |A \cap \Pi|. \tag{4.55}$$

Notice if Π is a Poisson process $N(A)$ for any A is a random variable and $N(A) \in \mathbb{N} \cup \{\infty\}$. We can now define [Kingman, 1993]

Definition 4.2.3 (Poisson Process)**.** *Suppose ν is a measure on a space $\mathcal{X}$. A Poisson process on $\mathcal{X}$ is a random subset of $\mathcal{X}$ such that if $N(A)$ is the number of points in the intersection of a measurable A and Π then (i) $N(A)$ is Poisson distributed random variable*

$$N(A) \sim \text{Poisson}\big(\nu(A)\big) \tag{4.56}$$

and (ii) if $A_1, \ldots, A_k$ is a collection of disjoint measurable subsets of $\mathcal{X}$ then the random variables $N(A_1), \ldots, N(A_k)$ are independent.

In the definition it is assumed if $\nu(A) = \infty$ then the Poisson random variable $N(A)$ is also fixed at ∞. It is easy to show

$$\mathbb{E}\big[N(A)\big] = \nu(A) \tag{4.57}$$

and for this reason ν is commonly denoted the *mean measure* of Π. An important point is that the Poisson process, in principle, allows an easy sampling scheme. Suppose $\nu(\mathcal{X}) = \infty$ and consider the following way to approximate a random sample D: (i) select a subset $A \subset \mathcal{X}$ where $0 < \nu(A) < \infty$. It follows $N(\Pi \cap A), N(\Pi \cap (\mathcal{X} \setminus A))$ are independent. (ii) sample $n \sim \text{Poisson}(N(A))$ and (iii) add n points sampled i.i.d. from $\nu(\cdot)/\nu(A)$. Repeat the construction for a new subset $A' \subset \mathcal{X} \setminus A$. The construction can be continued until a sufficiently large subset of $\mathcal{X}$ has been exhausted. Naturally, if $\nu(\mathcal{X}) < \infty$ one could simply select $A = \mathcal{X}$ to obtain an exact sample.

While the Poisson process is defined as a random subset of $\mathcal{X}$ and a CRM is a random measure on $\mathcal{X}$, the two are nevertheless closely related since the function $N(\cdot)$ is a random measure on $\mathcal{X}$ and fully characterizes the random set of the Poisson process. $N(\cdot)$ is commonly denoted a *Poisson random measure*. The distinction between finite additivity for the PP in eq. (4.56) and infinite for the CRM in eq. (4.47) can be overcome by a standard continuity argument [Kingman, 1993].

With these definitions in place we are now ready to state the main result. A CRM μ with Lévy-intensity ν can be represented as a linear function of a poisson random measure on $\mathbb{R}^+ \times \mathcal{X}$ with mean measure ν through [Kingman, 1967]:

$$\mu(A) = \int_{\mathbb{R}^+ \times A} sN(ds, dx). \tag{4.58}$$

or to put this in more familiar terms, by sampling the sequence $(w_i, \theta_i)_i \sim \text{PP}(\nu)$ and constructing μ through eq. (4.48) where $N \sim \text{PP}(\nu)$ denotes a Poisson process with mean measure ν, see also James [2005] for a comprehensive treatment on the construction of random measures from Poisson processes.

Completely random measures (and constructions based on these) can be characterized by the properties of ν. An important special case is if

$$\nu(ds, dx) = \rho(ds)\nu_0(dx) \tag{4.59}$$

in which case ν is said to be *homogeneous* and otherwise *non-homogeneous*. A NRMI based on a non-homogeneous Lévy intensity will have the special property of allowing statistical correlation between stick-length (group size of the induced partition) and spatial location of the sticks. However by far the most studied class of intensities are homogeneous intensities in particular the generalized gamma process [Brix, 1999] where:

$$\rho(s) = cs^{-a-1}e^{-bs}, b, c > 0, 0 < a < 1 \tag{4.60}$$

and by normalizing the induced NRMI one obtains the normalized generalized gamma process (NGG) [Lijoi et al., 2008].

A difficult problem for constructions based on CRMs is that the characterization of the posterior distribution which requires careful analysis. For a comprehensive reference on the available analytical tools see [James, 2002, James et al., 2009].

Priors based on non-homogeneous measures have a long history in the mathematical statistics community. They have primarily being applied for temporal survival data and planar spatial phenomena, see for instance [Ferguson, 1974, Hjort, 1990, Walker and Muliere, 1997]. In the recent years there has been some progress in tractable inference for non-homogeneous intensities due to [James et al., 2009, theorem 1] which also contain an application as well as [Lijoi and Prünster, 2010, chapter 3]. The recent work of Griffin and Walker [2011] contain additional details including discussion of a slice-sampling algorithm.

4.3 Random Graphs

This section considers relational data. By relational data refer to data which consists of relationships between discrete units. Examples could be a single set of units, *people*, and the relationship could be binary, *is-friends*. Other examples could be two sets of units, *people* and *books*, and the relationship could be *has-read*, indicating a particular person has read a particular book. This discussion offer several immediate generalizations. For instance we could consider multi-adic relationships such as the tri-adic relationship on the set

$$people \times people \times places \tag{4.61}$$

consisting of *someone-kissed-someone-at-someplace*. Notice this relationship is symmetric in the first two arguments, indicating if a kissed b under the bridge

then b kissed a under the bridge as well. Finally one could consider an additional complication by making the relationship more complicated than simply binary, in the example of books and people, the relationship considered could be *how-many-times-has-the-person-read-the-book*, in which case a natural choice for its value would be natural numbers $0, 1, \ldots$, rather than the Boolean label *has-read*. With fairly simple modifications, the ideas considered in this section applies to all cases. For simplicity we will therefore consider either dyadic relationships between two types of units, such as *has-read*, and when convenient relationships between objects of a single type, such as *is-friends*.

4.3.1 The Aldous-Hoover theorem

Any particular *observation* of relational data can be written as a d-dimensional array and so it is natural to study models of random d-dimensional arrays. We will later consider reasons not to consider this to be the only appropriate representation of relational data, however in the following it will be taken for granted. The discussion will follow that of random lists leading to the de Finetti theorem: First we will introduce appropriate notions of symmetry (for lists this was exchangeability) and later describe how this lead to a general characterization of the random array.

We will write (X_{ij}) for a random infinite 2-dimensional matrix

$$\begin{bmatrix} X_{11} & X_{12} & \cdots \\ X_{21} & X_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{4.62}$$

where for a realization (x_{ij}) of (X_{ij}) each element belong to a space $\mathcal{X}$. Such a random matrix is said to be either *separately* or *jointly* exchangeable if

$$\textit{(Seperately exchangeable:)} \qquad (X_{ij}) \overset{\mathrm{d}}{=} (X_{\sigma(i)\sigma'(j)}) \tag{4.63a}$$

$$\textit{(Jointly exchangeable:)} \qquad (X_{ij}) \overset{\mathrm{d}}{=} (X_{\sigma(i)\sigma(j)}) \tag{4.63b}$$

for any two permutation σ,σ'. Compare this definition to that of exchangeability, eq. (4.4). The Aldous-Hoover theorem, discovered independently by very different methods [Aldous, 1981, Hoover, 1979] (see also Austin [2008] for additional discussion of higher-dimensional arrays and further references) now states

Theorem 4.3.1 (Aldous-Hoover)**.** *A random array (X_{ij}) is seperately/jointly exchangeable if and only if it can be represented as a random function F :*

$[0,1]^3 \mapsto \mathcal{X}$ *such that*

$$\textit{(Seperately:)} \qquad (X_{ij}) \stackrel{\mathrm{d}}{=} (F(U_i, \tilde{U}_j, U_{ij})) \tag{4.64a}$$

$$\textit{(Jointly:)} \qquad (X_{ij}) \stackrel{\mathrm{d}}{=} (F(U_i, U_j, U_{ij})) \tag{4.64b}$$

for sets of variables $(U_i), (\tilde{U}_i), (U_{ij})$ *where* $U_i, \tilde{U}_i, U_{ij} \sim \mathrm{Uniform}([0,1])$ *iid.*

An important special case is when the graph is simple, that is, undirected meaning for all instances (x_{ij}) of (X_{ij}) it holds $(x_{ij}) = (x_{ji})$ that there are no self loops $x_{ii} = 0$ and the graph is binary $x_{ij} \in \mathcal{X} = \{0, 1\}$. In this case the Aldous-Hoover represented theorem can be recast as [Orbanz and Roy, 2013]

Theorem 4.3.2 (Aldous-Hoover for simple graphs)**.** *A random simple graph* (X_{ij}) *is jointly exchangeable if and only if it can be represented as a random symmetric function* $W : [0,1]^2 \mapsto [0,1]$ *(zero on the diagonal,* $W(x,x) = 0$*) such that*

$$(X_{ij}) \stackrel{\mathrm{d}}{=} \mathrm{Bernoulli}(W(U_i, U_j)) \tag{4.65}$$

for a list of iid. random variables (U_i)*,* $U_i \sim \mathrm{Uniform}([0,1])$*.*

The random function W is called the *graphon* and the above expression will later play an important role for describing network models.

When we discuss network models in chapter 6 we will refer extensively to the graphon. To give a specific example, recall the simple block-type model eq. (2.90) introduced in chapter 2. The definition was for a particular K:

$$z_i \sim \mathrm{Multinomial}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right) \tag{4.66a}$$

$$\theta_{\ell,k} \sim \mathrm{Beta}(b_1, b_2) \tag{4.66b}$$

$$A_{ij} \sim \mathrm{Bernoulli}(\theta_{z_i, z_j}). \tag{4.66c}$$

It is easy to see if W is the random function obtained by partitioning the unit interval into K intervals $V_1, \ldots, V_K$ such that:

$$V_i = \left[\frac{i-1}{K}, \frac{i}{K}\right[\tag{4.67}$$

then writing W as the random function obtained by first drawing $K(K+1)/2$ random $\theta_{\ell k}$ values according to eq. (4.66b) and setting

$$W(x,y) \equiv \sum_{\ell k} \theta_{\ell k} 1_{V_\ell}(x) 1_{V_k}(y) \tag{4.68}$$

(with the convention $\theta_{\ell k} = \theta_{k\ell}$) we obtain the corresponding random function. It takes little imagination to replace the finite set of intervals $(V_k)_{k=1}^K$ with that of a draw from the Poisson-Dirichlet process eq. (4.42); we will return to this construction later in chapter chapter 6.

4.3.1.1 Sparsity

An important consequence of the Aldous-Hoover theorem is the random network models are dense [Orbanz and Roy, 2013, Lloyd et al., 2013]. By dense we mean the number of edges scale as n^2. To be specific, for a random simple graph model (X_{ij}) we define the expected number of edges as

$$L(n) = \sum_{1 \leq i < j \leq n} x_{ij}. \tag{4.69}$$

If the graph is simply exchangeable it is a consequence of the Aldous-Hoover theorem 4.3.2 that

$$L(n) = \frac{W_0}{2} n(n-1), \qquad W_0 = \int_0^1 \int_0^1 du dv \; W(u,v). \tag{4.70}$$

Thus, any particular draw from the prior will either have no edges with probability 1 if $W_0 = 0$ or the number of edges will scale as n^2 and the same hold in expectation over W. It is widely believed the degree-distribution of many large networks follow a power-law[Newman, 2010, Newman et al., 2001, Strogatz, 2001]

$$p\big(\text{Fraction of vertices with } k \text{ edges}\big) \sim k^{-\gamma}, \qquad \gamma > 0 \tag{4.71}$$

where γ roughly falls in the interval $2 - 3$. In this case the number of edges grow as $\mathcal{O}(n^\alpha)$, $\alpha < 2$ [de Solla Price, 1965, Barabási and Albert, 1999]. Such as statement is of course subject to many qualifications. After all, we only have access to a single network and not an infinite-dimensional ensemble wherein we can take the limit eq. (4.69), and even if we registered consecutively larger parts of a social network one should worry if the corresponding density of friendships was really reflecting the way the social network was being sampled. However it seems to be a plausible assumption that in many settings the units corresponding to vertices has physical limitations, for instance and upper bound on the number of friendships, that makes it reasonable to say the number of edges scale slower than $\mathcal{O}(n^2)$.

While this scaling behavior seems unobtainable for exchangeable random graphs, we will briefly mention some recent work that imply such as negative lesson should be subject to qualifications. Let's for a moment not consider networks

but something quite different, namely a large shallow tank of water and very long sides, say A meters. Assume the water in the tank is polluted with small, non-interacting dark particles of pollen lying on the surface of the water. We assume the concentration of particles is 1000 particles per liter of water and above the tank we assume there is a camera pointing directly downwards. The location of the camera is fixed and we assume the aperture of the camera is so small it can only see a small rectangular region in the middle of the tank of side length a.

We assume the following setup: At different points in time we (1) stir the tank and wait for the water to settle down (2) take a picture with the camera. Each picture of the camera will contain a number M of particles and, since it is a digital camera, the particles will have (x, y) positions relative to the aperture of the camera, $L = (x_i, y_i)_{i=1}^M \subset \mathbb{R}^2$. Clearly repeating the procedure of stirring and taking snapshots will induce a distribution over discrete subsets L of $\mathbb{R}^2$. It is not hard to see M will be binomially distributed and, if the tank is very large, very well approximated by a Poisson distribution. Furthermore it is not hard to see the distribution of L (for a particular camera) is the same as if we could "freeze" the water in the tank, cut it into small cubes $w_i \times w_j$, $w_i = [w_i^a, w_i^b[$ and apply a permutation σ on the cubes to produce a new tank $w_{\sigma(i)} \times w_{\sigma(j)}$.

Suppose each set L is given the following interpretation: The coordinate (x_i, x_j) corresponds to assuming there are two vertices i and j and they are connected by an edge (ij). Since $x_i \neq x_j$ with probability 1 for different i, j this network would (with probability 1) contain M edges and $2M$ vertices. If we made the camera zoom out to capture 4 times the area, the new (induced) network $L' \subset \mathbb{R}$ could be expected to contain 4 times as many vertices and 4 times as many edges on average; however notice the distribution of the subset $L = L' \cap ([0, a[\times[0, a[)$ would be the same. It would appear the induced networks both shared many of the important features of exchangeability while the expected number of edges only grew proportionally to the number of vertices. An issue with this argument is the network is somewhat trivial and only containing disconnected edges, however if we assumed the available choices of x and y was limited, for instance because the camera only captured a finite number of pixels, some of the positions would coincide and the resulting graph would be non-trivial. Such a construction was proposed by Caron [2012] (see also Caron and Fox [2014]) and may be briefly sketched as the following generative model

$$\mu \sim \mathrm{CRM}(\nu) \tag{4.72a}$$

$$L \sim \mathrm{PP}(\mu \times \mu) \tag{4.72b}$$

where the CRM and Poisson process discussed in section 4.3.1.1. Notice the result of this procedure, L, is a a random set of points on the space where ν (the intensity of the CRM, see eq. (4.50)) is defined. For additional discussion

see [Caron, 2012, Caron and Fox, 2014]. Different choices of ν will control the properties of L and allow the construction of random sparse networks [Caron and Fox, 2014]. It might appear the above construction, a model of random graphs that are not dense, is in conflict with the Aldous-Hoover theorem, however this is not the case because the construction is *not* a model of exchangeable arrays in the sense of eq. (4.63) but of points sets obeying a different type of invariance [Kallenberg, 2005, Chapter 9]; we expect this construction to play an important role in the coming years and will briefly discuss it later in chapter 6, however since it has played no further role in our work we will not provide additional details here.

4.4 Random Hierarchies

The last section will treat priors over *rooted trees* also known as *hierarchies*. Since a tree is a graph, it might be tempting to consider a theory of random trees as closely related to that of random graphs, however the treatment will be much more closely related to that of a partition.

Algorithm 1 Generate a fragmentation $t \leftarrow \text{FRAG}(B)$

```
t ← {B}
if |B| ≥ 2 then
  π ← CRP(B, α)
  for b_k ∈ π do
    t ← t ∪ FRAG(B)
  end for
end if
return t
```

The central object we will study is a *fragmentation*. A fragmentation can be thought of as starting with a set B and then (i) construct a partition π of B (ii) recursively apply the partition scheme used in (i) to each block $B_k \in \pi$ (iii) terminate when one is left with singletons. Fragmentations are thus simplest described by an algorithm which generate partitions. Recall $\pi \sim \text{CRP}(B, \alpha)$ denotes that π is a partition of B obtained from a Chinese restaurant process with parameter α. For instance if $B = \{1, 4, 5, 8\}$ then it may be the case $\pi = \{\{1, 4\}, \{5\}, \{8\}\}$. A simple algorithm for generating fragmentations is then illustrated in algorithm 1 and a fragmentation of B could be t_B where

$$t_B \equiv \text{FRAG}(B) = \{\{1, 4, 5, 8\}, \{1, 4\}, \{5, 8\}, \{5\}, \{8\}, \{1\}, \{4\}\}. \tag{4.73}$$

To give a formal definition: a fragmentation t_B of B is a collection of non-empty subsets of B such that: (i) $B \in t_B$ and (ii) if $|B| \geq 2$ there is a partitioning

$\pi^B = \{B_1, \ldots, B_K\}$ of B such that $K \geq 2$ and

$$t_B = \{B\} \cup t_{B_1} \cup t_{B_2} \cdots \cup t_{B_K} \tag{4.74}$$

where each t_{B_k} is a fragmentation of B_k. Notice it follows from the definition t_B must contain all singletons: $\{i\} \in t_B$ if and only if $i \in B$.

Fragmentations are easily seen to be equivalent to rooted trees. Each element of t_B is identified with a vertex in the tree such that $\{B\}$ is the root of the tree and each singleton $\{i\}, i \in B$ is a leaf. Edges are induced by the definition eq. (4.74): In the used notation we would identify vertex corresponding to $\{B\}$ as having K children, $B_1, \ldots, B_K$ and we would say each tree t_{B_i} is a *subtree* of t_B. Notice that according to this definition no vertex in the tree can have a single child.

Fragmentations, and trees in general, are objects of great practical importance. The earliest pioneering studies was as pure combinatorics and go back to Borchardt [1860], Schröder [1870], Cayley [1889], however as statistics began to play a greater role in the study of evolution so was hierarchies naturally made an object of study from a statistical perspective as a model of cladograms. For modern treatments see Aldous [1996], Bertoin [2001], Haas et al. [2008] and McCullagh, Pitman, Winkel, et al. [2008], the later serving as the primary reference of this section.

What we are interested in is models of random fragmentations. That is, simply a probability distribution over the set of all fragmentations of finite sets of all sizes. This is analogous to our desire to find a distribution over all partitions of all sets of finite size.

The project proceeds entirely parallel to that of partitions and arrays: First we identify (postulate) an appropriate symmetry condition (for partitions this was exchangeability, for arrays this was joint or separate exchangeability), then this symmetry induces an appropriate characterization in terms of an infinite-dimensional object. For exchangeable sequences this was a random measure (for partitions the paint-box characterization) and for exchangeable arrays it was the Aldous-Hoover representation, specifically the graphon for joint exchangeable graphs. Finally this object can be used to specify a model for the discrete object in question that fulfills the desired invariance, an example was the one and two parameter Chinese restaurant process for partitions.

Where our *presentation* will depart slightly from this scheme is the notion of exchangeability most appropriate for fragmentations will be more restrictive. This will significantly limit the family of possible exchangeable random fragmentation models. We will in this presentation not provide the underlying (infinite-dimensional) characterization corresponding to the paint-box or graphon and

only describe the distribution over finite trees analogous to the Chinese restaurant process. We hope this omission does not obscure the similarities to the preceding sections we encourage the reader to consult [McCullagh et al., 2008] for the details.

4.4.1 Exchangeable fragmentations

As for the case of the robot in the introduction to this chapter, an important role of exchangeability is ensuring models of different sizes are consistent under marginalization. To make the notation less cumbersome we will define the notion in terms of *finite* trees. Consider first a partition π of B for some finite B. Assume A is some set and $A \cap B \neq \emptyset$. Recall the projection operation of π onto $A \cap B$ from eq. (4.18) was defined as

$$\mathrm{proj}_A(\pi) \equiv \{b \cap A : b \in \pi, b \cap A \neq \emptyset\}\,. \tag{4.75}$$

Notice the projection operator also work for trees. Specifically if t_B is a fragmentation of B and $A \cap B \neq \emptyset$ then $t_{A\cap B} = \mathrm{proj}_A(t_B)$ is also a fragmentation. In a similar fashion to the theory of exchangeable partitions section 4.2, denote by T_B a *random fragmentation* of the set B. A model of random fragmentations is then a collection of all such random variables $(T_{[n]})$.

We are now in a position to state the symmetry for fragmentation. We will say a model of random fragmentation is

- *Consistent* if, for all $A \subset B$ such that $A \neq \emptyset$, the projection $\mathrm{proj}_A(T_B)$ is distributed as $T_{A\cap B}$.
- *Markovian* if, letting $\Pi_B = \{B_1, \ldots, B_K\}$ denote the stochastic variable corresponding to the split of B at the root, the K trees

  $$\mathrm{proj}_{B_1} T_B, \ldots, \mathrm{proj}_{B_K} T_B \tag{4.76}$$

  conditional on Π_B are i.i.d. distributed as $T_{B_1}, \ldots, T_{B_K}$.

A random model of fragmentations with both of these properties will be called exchangeable. Notice the Markovian property is implicit in a generative model such as algorithm 1 and naturally arise if we do not wish subtrees to influence each other, meanwhile the consistency property, while less obvious from a generative perspective, is analogous to that of eq. (4.19) for partitions.

Remarkably, the above two requirements along with a further requirement discussed below are enough to narrowly identify the possible distributions. We

will only outline the argument which can be found in [McCullagh et al., 2008]. Firstly, an exchangeable model of random fragmentations can be characterized by its *splitting rule*, the equivalent of the EPPF for partitions. Assuming t_B fragment B into $\pi = \{B_1, \ldots, B_K\}$ then the splitting rule is the function s:

$$P(\Pi_B = \pi) = s(n_1, n_2, \ldots, n_K) \tag{4.77}$$

where $n_k = |B_K|$ and $s(1) = 1$. Naturally the function s is symmetric in its coordinates. By the Markovian property the full distribution of t_B is then of the form:

$$P(T_B = t) = \prod_{B \in t} s(|B_1|, |B_2|, \ldots, |B_K|) \tag{4.78}$$

where it is implied t fragments B into the sets $B_1, \ldots, B_K$. If we assume the splitting rule s is of the *Gibbs form*:

$$s(n_1, n_2, \ldots, n_K) = \frac{a(K)}{c(n)} \prod_{k=1}^{K} w(n_k), \qquad n = \sum_{k=1}^{K} n_k \tag{4.79}$$

for some $w(n) \geq 0$ if $n \geq 1$, $c(n) > 0$ if $n \geq 2$ and $a(k) \geq 0$ $k \geq 2$ with the special values: $w(1) = a(2) = 1$ we can now state the main result due to McCullagh, Pitman, Winkel, et al. [2008] and is here repeated nearly verbatim.

Theorem 4.4.1 (McCullagh-Pitman-Winkel)**.** *Let s be the splitting rule for random fragmentation model P. If s is of the Gibbs form eq.* (4.79) *and consistent, then s is associated with the two-parameter Ewens-Pitman family*

$$w(n) = \frac{\Gamma(n-d)}{\Gamma(1-d)},\ n \geq 1 \quad \text{and} \quad a(k) = d^{k-2}\frac{\Gamma(k+\alpha/d)}{\Gamma(2+\alpha/d)},\ k \geq 2. \tag{4.80}$$

including limits in d. $c(n)$ is determined by normalization. The allowed parameter range is as follows

$0 \leq d < 1$ **and** $\alpha > -2d$**:** *Multifurcating with arbitrary block numbers*

$d < 0$ **and** $\alpha = -md$**:** *Multifurcating with no more than m blocks where m is an integer*

$d < 1$ **and** $\alpha = -2d$**:** *Binary hierarchies*

$d = -\infty$ **and** $\alpha = m$**:** *for integer $m \geq 2$. The "recursive coupon collector" of McCullagh et al. [2008]*

$d = 1$**:** *Singleton blocks.*

To apply Gibbs a fragmentation trees to a particular model one need to compute the posterior likelihood of any particular tree which require one to compute $c(n)$ in the above. We will focus on the first range of parameters, $0 \leq d < 1$ and $\alpha > 0$. Using the definition eq. (4.78) and the form of the splitting rule eq. (4.79) it is simple to verify $s(n_1, \ldots, n_K)$ must be of the form

$$s(n_1, \ldots, n_K) = \frac{\left(\frac{\alpha}{d}\right)^{(K)} d^{K-1}}{\alpha^{(n)} - (-d)^{(n)}} \prod_{k=1}^{K} (-d)^{(n_k)}. \tag{4.81}$$

where again $x^{(k)} = \frac{\Gamma(k+x)}{\Gamma(1+x)}$. This is simply the density of the two-parameter Chinese restaurant process of eq. (4.44) conditional on there being at least two blocks.

4.5 Discussion

It is interesting to relate the discussion of this chapter to that of the previous two. Firstly, Bayesian non-parametrics, as a research subject, is perhaps best understood as finding (i) relevant types of data (ii) relevant symmetries and invariance principles appropriate for that type of data. Having arrived at a good question under (i)–(ii), it is a mathematical problem to derive representations of the probability distribution that satisfies both. Machine learning is naturally interested in the same problems, however with a strong preference for ease (or often simply feasibility) of implementation at the cost of generality.

This asymmetry means the mathematical statistics literature has historically been ahead of machine learning. A particular object (a list, an array, a fragmentation, etc.) is very often first studied in mathematical statistics and then, sometimes decades later, the same type of object is appear and is studied in the machine learning literature in the context of particular problems. This provides ample reasons to search the mathematical literature for novel objects or results, and consult the machine-learning literature for applications and problems. This, and the fact the mathematical statistics field is relatively mature, implies references older than the past few years still have immense practical value, a fact I only discovered relatively late.

This chapter completes the theoretical building-blocks. The next two chapters will, respectively, treat the practical problem of inference in discrete model and applying the previous material to network modelling.

Chapter 5

Inference

In this chapter, we will consider the problem of inference. To illustrate the problem, suppose we have a Bayesian model consisting of a large collection of parameters x. The variables may be continuous or discrete, and they may encode a complicated data structure such as a hierarchy. For simplicity and definiteness, the reader is invited to consider the case where x consists of a list of variables, $x \equiv (x_i)_{i=1}^n$ and where each x_i is discrete: $x_i \in \mathcal{X}_i$.

In a Bayesian framework we will typically assume we have access to some data Y (which may also be of any sort) and a model

$$p(Y, x) = p(Y|x)p(x). \tag{5.1}$$

Suppose we observe data Y and assume the model is all we know. In this case all the comments in chapter 3 do not apply and we are left with Bayes theorem. The posterior density becomes:

$$p(x|Y) = \frac{p(Y|x)p(x)}{\int dx'\ p(Y|x')p(x')}. \tag{5.2}$$

To appropriately formulate the main results we will continue to distinguishing between probability measure and densities as in chapter 4, however we will follow the custom in the literature of using Greek letters for the probability density and in particular use π for the probability of interest, that is, corresponding to

the density $p(x|Y)$ in eq. (5.2). In addition we will often use the same symbol for the distribution and density. Assuming μ is the standard (Borel) measure on $\mathcal{X}$ then we will write

$$\pi(dx) = \pi(x)\mu(dx), \quad dx \subset \mathcal{X} \tag{5.3}$$

where $\pi(x) = p(x|Y)$ in the notation of eq. (5.2). We will otherwise continue to work with the usual notation of measure and integration theory, see Kallenberg [2002, chapter 1] for a comprehensive reference. In itself, formulating a theory for the (correct) probability distribution π may seem quite limited since we can normally only compute $p(x, Y)$ in closed form, nevertheless it is customary to work with the normalized distribution π since the derived results will be unaffected by the scaling.

5.1 The inference problem

What we have not concerned ourself with so far is what we are supposed to do with the posterior distribution π. In this chapter, we will assume only one thing concerns us: The computation of expectations. More exactly evaluation of expectations of the form:

$$\mathbb{E}[f] = \int \pi(dx)\ f(x) \tag{5.4}$$

for one or more measurable functions f. We will assume f is well-behaved and the integral always converges. This type of averages arises in many circumstances. Suppose for instance in additional to the data Y there is some additional unobserved data Y_u and x capture all latent information about the observations: $p(Y, Y_u|x) = p(Y|x)p(Y_u|x)$. In this case $p(Y_u|Y) = \int dx\ p(Y_u, x|Y) = \int dx\ p(Y_u|x)p(x|Y)$ thus:

$$p(Y_u|Y) = \mathbb{E}[f], \qquad f(x) = p(Y_u|x). \tag{5.5}$$

Alternative situations could be that $f(x)$ represents a cost or reward of a state x, a physical quantity depending on x such as energy, force or position or the indicator function on an set A: $f(x) = 1_A(x)$. This chapter treats the problem of evaluating (or rather, approximating) such integrals using Monte Carlo methods. Monte Carlo methods denotes a very large class of algorithms, methods and theoretical results and Monte Carlo methods are in turn a subset of all inference methods applicable to Bayesian models. We will not attempt to review this literature nor will we give a particular throughout account of Monte Carlo methods. Rather, after a brief introduction, the chapter will be focused

rather narrowly on the case where the variables of interest x corresponds to a partition.

A reader who is interested in Monte Carlo methods in general, especially for models with continuous parameter spaces, should consult some of the many excellent general references on the topic [Neal, 1993, Doucet, 2001, Gilks, 2005, Robert and Casella, 1999, Rubinstein and Kroese, 2011]; In addition, a comprehensive theoretical perspective on Markov chains is found in Kallenberg [2002].

5.2 Monte Carlo methods

In most situations integrals such as eq. (5.4) will be analytically intractable due to the form of $p(x|Y)$. In this case numerical integration can be attempted, see for instance Press et al. [1990], Davis and Rabinowitz [2007]. However, if the system contains many non-trivial dimensions, these methods will be insufficient. It is in this regime Monte Carlo methods becomes relevant.

Monte Carlo methods rest upon a very simple idea. Assume π is the probability density of interest for a random variable X in $\mathcal{X}$. The strong law of large numbers [Scheaffer and Young, 2009] states that, assuming $f(X)$ has finite variance and letting $(X_t)_{t=1}^n$ be n realizations of X, then almost surely

$$\frac{1}{n}\sum_{t=1}^{n} f(X_t) \to \mathbb{E}[f] \quad \text{as} \quad n \to \infty \tag{5.6}$$

In other words, suppose we can generate n samples $x_1, \ldots, x_n$ from π then the average $(f(x_1) + \cdots + f(x_n))/n$ will approximate $\mathbb{E}[f]$. This method is known as the Monte Carlo method[Metropolis and Ulam, 1949]. The method may be extended in two important ways. Suppose we cannot *sample* from π, however we can *evaluate* π. In this case we can consider an alternative method. Suppose we can sample from another distribution t defined over $\mathcal{X}$ with decomposition $t(dx) = t(x)\mu(dx)$ and such that $t(x) > 0$ if $\pi(x) > 0$. In this case we can rewrite the problem eq. (5.4)

$$\mathbb{E}_\pi[f] = \int \pi(dx)\ f(x) = \int \mu(dx)\ t(x)\frac{f(x)\pi(x)}{t(x)} = \mathbb{E}_t\left[\frac{f\pi}{t}\right]. \tag{5.7}$$

Which suggests evaluating the expectation by generating n samples $x_1, \ldots, x_n$ from $t(\cdot)$ and make use of the approximation

$$\mathbb{E}_\pi[f] \approx \sum_{t=1}^{n} \frac{f(x_t)\pi(x_t)}{t(x_t)}. \tag{5.8}$$

This method is known as *importance sampling* [Hammersley and Handscomb, 1964]. This method is less general than the derivation might suggest. Firstly, while most functions f of interest has finite variance, this is far less certain of the ratio $f\pi/t$ and secondly, even if this expression is convergent the variance will often be so high as to be problematic in practice. Standard results[Robert and Casella, 1999, Rubinstein and Kroese, 2011] show the t which minimize the variance of $\mathbb{E}_\pi[f] = \mathbb{E}_t[f\pi/t]$ is given by

$$t_{\text{opt}}(x) = \frac{|f(x)|\pi(x)}{\int dx' \ |f(x')|\pi(x')} \tag{5.9}$$

Which, assuming f is more or less uniform, states the unsurprising result t should approximate π. Since π was assumed hard to sample from to begin with this is often unfeasible to obtain in practice.

The second extension is a standard trick: whenever one has a stochastic algorithm which relies on i.i.d. random samples from some distribution, and those samples are non-trivial to obtain, one should consider replacing the i.i.d. samples by correlated samples. This is the topic of the next section.

5.3 Markov Chain Monte Carlo

The idea of Markov Chain Monte Carlo methods can be simply stated as replacing the independent copies of X, $(X_t)_{t=1}^n$, in eq. (5.4) by dependent variables $X_1, \ldots, X_n$. The definitions and results in the following section can all be found in Tierney [1994]. A particular simple way to define dependent variables is using a Markov Chain. A Markov chain is a sequence of random variables $(X_t)_{t=0}^n$ with the property that given a particular state X_i takes a value x_i, the past $(X_t)_{t=0}^{i-1}$ and future $(X_t)_{t=i+1}^n$ states are independent. This implies for all t:

$$P(X_t \in A | X_0 = x_0, \cdots, X_{t-1} = x_{t-1}) = P(X_t \in A | X_{t-1} = x_{t-1}) \tag{5.10}$$

for any measurable set A. The conditional distribution on the right-hand side plays a crucial role and is known as the *transition kernel*.

Formally, a transition kernel is a function T on $\mathcal{X} \times \mathcal{B} \mapsto [0, 1]$ where $\mathcal{B}$ is the measurable sets of $\mathcal{X}$ such that (i) $T(x, \cdot)$ is a probability measure for all $x \in \mathcal{X}$ and (ii) $T(\cdot, A)$ is a measurable function for all $A \subset X$. In this notation, the Markov condition eq. (5.10) can be written

$$P(X_t \in A | X_0 = x_0, \cdots, X_{t-1} = x_{t-1}) = T_{t-1}(x_{t-1}, A). \tag{5.11}$$

Observe with this definition we can define the two-step transition kernel $T_{t-2}^{(2)}(x_{t-2}, A)$ by applying eq. (5.11) twice: [Tierney, 1994]

$$\begin{aligned}
&P(X_t \in A | X_{t-2} = x_{t-2}) \\
&= \int_{x_{t-1} \in \mathcal{X}} P(X_t \in A | X_{t-1} = x_{t-1}) P(X_{t-1} \in dx_{t-1} | X_{t-2} = x_{t-2}) \\
&= \int_{x_{t-1} \in \mathcal{X}} T_{t-1}(x_{t-1}, A) T_{t-2}(x_{t-2}, dx_{t-1}) \\
&\equiv T_{t-2}^{(2)}(x_{t-2}, A).
\end{aligned} \tag{5.12}$$

This inspires the following general definition for all $t \geq s \geq 1$ (the product is understood as denoting n integral operators)

$$\begin{aligned}
&P(X_t \in A | X_{t-s} = x_{t-s}) \\
&= \left(\prod_{i=1}^{s-1} \int_{x_{t-i} \in \mathcal{X}} \right) T_{t-1}(x_{t-1}, A) \prod_{i=1}^{s-1} T_{t-i-1}(x_{t-i-1}, dx_{t-i}) \\
&\equiv T_{t-s}^{(s)}(x_{t-s}, A).
\end{aligned} \tag{5.13}$$

With this definition in mind we can define the marginal distribution $P(X_t \in A)$ at any time t given some initial value x_0 of X_0

$$\pi^{(t)}(A) \equiv P(X_t \in A | x_0) = T_0^{(t)}(x_0, A). \tag{5.14}$$

Furthermore, using the definition eq. (5.13) it follows π obeys the following relationship:

$$\pi^{(t)}(A) = \int \pi^{(t-1)}(dx_{t-1}) \; T_{t-1}(x_{t-1}, A). \tag{5.15}$$

Where we have included a time index on the transition kernels T_t. If the transition kernels are independent of time, that is, $T_t(x, A) = T_0(x, A)$ for all x, A, t the Markov chain (and equivalently, the transition kernel) is called *homogeneous*. Otherwise it is called *non-homogeneous*. While we will later consider non-homogeneous chains, for now we will assume the Markov chains are homogeneous unless otherwise mentioned. The result eq. (5.15) inspires the following definition:

For a transition kernel T we say a distribution π is *invariant* (or *stationary*) provided [Tierney, 1994]

$$\pi(A) = \int \pi(dx) \; T(x, A) \tag{5.16}$$

for all measurable $A \subseteq \mathcal{X}$. Notice the definition trivially extends to general measures μ provided $\mu(\mathcal{X}) < \infty$.

Suppose π, the distribution we wish to obtain samples from, is the stationary distribution of some transition kernel T. Then, given $x_0 \sim \pi$, we can generate n samples from π, $(x_t)_{t=1}^n$, by sampling x_t from $T(x_{t-1}, \cdot)$ for $t = 1, \ldots, n$ and use these samples in the approximation eq. (5.4).

There are two issues with this scheme. The first is we cannot assume the initial point x_0 is distributed as π, the second is the samples will not be i.i.d. samples from π. Convergence results for Markov chains are thus concerned with showing which conditions must apply for the transition kernel for allowing this scheme to converge. Before this, however, we will illuminate the balance condition eq. (5.16) from an intuitive perspective.

5.3.1 The balance condition

Consider the balance condition eq. (5.16) in the case where $\mathcal{X}$ is finite, $\mathcal{X} = \{x^1, \ldots, x^M\}$, and the kernel is homogeneous. In this case for all $x \in \mathcal{X}$:

$$\pi(x) = \sum_{k=1}^{M} T(x|x^k)\pi(x^k). \tag{5.17}$$

The balance condition can now be given a simple interpretation. Suppose each state x^k of $\mathcal{X}$ correspond to a barrel and $\pi(x^k)$ is the amount of water (in liters) in the barrel. Drawing a random sample from π is now equivalent to selecting a barrel x^k with probability proportional to the amount of water in the barrel. This can be done as follows: First, put a mark on the side of the barrel indicating the amount of water. Next, empty all the barrels into a large basin. Third, place a microscopic radio transmitter of neutral buoyancy in the basin and stir. Fourth, pour the water back in the barrels to their original level. The barrel x^k with the radio transmitter has then been selected with probability $\pi(x^k)$.

Suppose we come up with the following time-saving scheme: Instead of filling and refilling the barrels, we connected the barrels with pipes such that for any two barrels there are two pipes connecting them. To each pipe there is a pump, and the rate of flow of all pumps is carefully adjusted such that the amount of water in each barrel do not change, for instance by having the same flow from barrel A to B as from B to A. If this system is left to its own for sufficiently long time the flow will eventually have re-distributed the water evenly, imitating the effect of collecting-and-redistribution scheme. Accordingly, the location of the radio transmitter will again be random and the particular barrel it is located in will be randomly sampled with probability π. The balance equation eq. (5.17) can be framed as the pumps keeping the water levels constant and the image

of the current state x_t, corresponding to the radio transmitter, as flowing in a state space provide a useful illustration of the conditions for convergence.

5.3.2 Convergence

In this section, we will briefly review some basic convergence results. To discuss these requires introducing properties of a Markov chain useful for expressing the convergence conditions. The definitions are somewhat technical since uncountable spaces $\mathcal{X}$ introduces measure-theoretical difficulties and for this reason it is worth having the illustration with the barrels, pipes and the radiotransmitter in mind to illustrate which situations the definition are trying to avoid.

Consider a Markov chain $(X_t)_{t=0}^{\infty}$ with invariant (in the sense of eq. (5.16)) measure π and transition kernel $T(\cdot, \cdot)$ on a space $\mathcal{X}$ with σ-algebra $\mathcal{F}$. We wish to avoid that the Markov chain behaves fully deterministically or that it will not visit some region of state space infinitely often. The following definitions largely follow Tierney [1994] with slightly more explicit notation. For any measurable A define the random variable τ_A through

$$\tau_A = \inf\{t \geq 1 : X_t \in A\} \tag{5.18}$$

with $\tau_A = \infty$ if the chain never visit A. τ_A denote the first time the chain enters a region A. Clearly a chain which, with some probability, never visits a "large" region A will (with the same probability) give samples that are uninformative regarding that region. Conversely, it should not matter if a chain fails to visit a single point with measure 0. After all, for an uncountable space the chain is guaranteed to only visit a countable subset. The following definitions will be important

- The chain is said to be *ϕ-irreducible* if there exist a σ-finite measure ϕ on $\mathcal{X}$ such that $\phi(\mathcal{X}) > 0$ and for any initial state $x_0 \in \mathcal{X}$ and any measurable $A \in \mathcal{F}$ such that $\phi(A) > 0$ then $P(\tau_A < \infty | X_0 = x_0) > 0$. For our purpose it will suffice to take $\phi = \pi$.
- *Aperiodicity* is the simple requirement the chain does not jump between states in a deterministic manner. Formally, a π-irreducible Markov chain is said to be *aperiodic* if there exist no partition $B_0, \ldots, B_{k-1}$ where $k \geq 2$ such that $x_0 \in B_0$, $\pi(B_k) > 0$ for all k and for all t:

$$T^{(t)}(x_0, B_{t \pmod k}) = 1. \tag{5.19}$$

- *Recurrence* is the property there is no states the chain will only visit a finite number of times. Formally, we say a π-irreducible Markov chain

with stationary distribution π is *recurrent* if for all $A \subset \mathcal{X}$ such that $\pi(A) > 0$

$$P(\tau_A < \infty | X_0 = x_0) > 0 \text{ for all } x_0 \in \mathcal{X} \tag{5.20}$$

$$\text{and } P(\tau_A < \infty | X_0 = x_0) = 1 \text{ for } \pi\text{-almost all } x_0 \in \mathcal{X}. \tag{5.21}$$

- As a corollary to the above, the chain is said to be *Harris recurrent* if $P(\tau_A < \infty | X_0 = x_0) = 1$ for all $x_0 \in \mathcal{X}$.
- Finally the chain is said to be *ergodic* if it is positive Harris recurrent and aperiodic.

The following result shows the above four properties are in a sense necessary and sufficient for convergence.

Theorem 5.3.1. *Suppose a Markov chain with transition kernel T is π-irreducible and admits π as the stationary distribution. Then T is positive recurrent and π is the unique invariant distribution of T. If T is also aperiodic, then, for π-almost all $x_0 \in \mathcal{X}$*

$$\|T^n(x_0, \cdot) - \pi(\cdot\ )\|_{var} \to 0 \text{ for } n \to \infty \tag{5.22}$$

If T is Harris recurrent, then the convergence occurs for all $x_0 \in \mathcal{X}$.

In the result eq. (5.22) the variational distance between two measure μ, ν is defined as

$$\|\mu - \nu\|_{\text{var}} = \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| - \inf_{A \subseteq X} |\mu(A) - \nu(A)| \tag{5.23}$$

where the supremum and infimum is taken over all measurable A.

The formulation of the above result follows Tierney [1994]. For proof and additional details see Nummelin [1984], Athreya, Doss, and Sethuraman [1992], Rosenthal [2001]. The above conditions are also necessary if the relationship eq. (5.22) hold for *all* x Tierney [1994]. Keep in mind that theorem 5.3.1, taken alone, only establish we will eventually converge towards a single sample, not that the average eq. (5.6) converges. This is establish in the following result

Theorem 5.3.2. *Suppose $(X_t)_{t=0}^{\infty}$ is ergodic with equilibrium distribution π and suppose f is real-valued and $\int d\pi\ |f| < \infty$. Then for any initial distribution of X_0*

$$\frac{1}{n} \sum_{t=1}^{n} f(X_t) \to \int \pi(dx) f(x) \quad \text{for } n \to \infty \text{ (almost surely).} \tag{5.24}$$

The formulation is again taken from Tierney [1994], for proof see Revuz [1975, theorem 3.6].

On one hand, the conditions for the above results are very minimal and insofar one relies on standard constructions (see the next section) they are unlikely to be violated. On the other hand they do not say how *fast* the chain (or the average eq. (5.6)) will converge. There exist a number of convergence result which all impose additional constraints on the transition kernel [Tierney, 1994, Nummelin, 1984, Chan, 1989, Rosenthal, 1995a]. These results either come with conditions that are hard to verify for any particular transition kernel or, alternatively, contain constants in the results which cannot be estimated in practice. For this reason, while nearly any proposed Markov chain based method can be expected to converge in the sense of theorems 5.3.1 and 5.3.2, it may do so very slowly and the question how well a particular sampler (by which we mean a transition kernel) actually works is by a large empirical.

5.4 Constructing samplers

In this section, we review two standard techniques for constructing transitions kernels which obey the invariance condition eq. (5.16).

5.4.1 Gibbs sampling

Gibbs sampling, while arguably not as general as Metropolis-Hastings sampling which we will consider in the next section, is nevertheless one of the most common single method for Bayesian inference by Markov chain Monte Carlo and for all problems that admit a computationally feasible implementation it is considered the go-to method.

The basic idea behind Gibbs sampling is the following. Suppose each element $x \in \mathcal{X}$ is represented as a product of d spaces, ie. $x = (x_1, \ldots, x_d)$ and $\mathcal{X} \equiv \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Write x_i for the variables belonging to subspace i, ie. $x_i \in \mathcal{X}_i$ and correspondingly $x_{\setminus i}$ for the other variables:

$$x_{\setminus i} \in \mathcal{X}_1 \times, \cdots, \mathcal{X}_{i-1} \times \mathcal{X}_{i+1} \times \cdots \mathcal{X}_d \tag{5.25}$$

In a similar vein we will introduce stochastic variables X_i and $X_{\setminus i}$ and let

$$\pi(X_i \in \cdot | X_{\setminus i} = x_{\setminus i}) \tag{5.26}$$

denote the conditional distribution of $X_i|X_{\setminus i}$. Introducing time-indices t to arrive at $x_{t,i}, x_{t,\setminus i}$ (and let x_t denote the full state), the Gibbs sampler generates the next state of the chain x_{t+1} from x_t by first setting $x_{t+1} = x_t$ and then updating each subspace using the conditional probability eq. (5.26):

- Iterate $i = 1, \ldots, d$:
  - Generate: $x_i^* \sim \pi(\cdot|X_{\setminus i} = x_{t+1,\setminus i})$
  - Update: $x_{t+1} = (x_{t+1,1}, \ldots, x_{t+1,i-1}, x_i^*, x_{t+1,i+1}, \ldots, x_{t+1,d})$.

There is nothing special about the order in which the variables are iterated over. The stationarity condition eq. (5.16) is straight-forward to verify through some tedious algebra.

Gibbs sampling was originally invented in statistical physics in the context of Ising spin systems by Ehrman, Fosdick, and Handscomb [1960] (Ehrman et al. [1960] in turn claim the idea is closely related to Metropolis et al. [1953], Wood and Parker [1957], Fosdick [1957], however the two former references fall within the Metropolis-Hastings framework and the third reference was not obtainable).

The first application of Gibbs sampling in machine learning was in image processing [Geman and Geman, 1984] and was later popularized in a number of important publications, c.f. [Tanner and Wong, 1987, Gelfand and Smith, 1990, Casella and George, 1992, Liu, 2008]. Gibbs sampling has been subject to a number of important modifications suitable for different conditions. This include *blocking and collapsing* [Liu, 2008, 1994], *data augmentation* [Tanner and Wong, 1987, Van Dyk and Meng, 2001] and general *axillary variable techniques* [Higdon, 1998] such as *slice sampling* [Neal, 2003]. A good overview can be found in [Brooks et al., 2011].

As mentioned in the introduction, Gibbs sampling is often considered the default method when at least *some* of the marginal distribution has an tractable analytical form which admits easy sampling, and even when this is not the case Gibbs-inspired techniques such as metropolis-within-Gibbs[Gilks et al., 1995, Tierney, 1994] can sometimes be applied.

5.4.2 Metropolis-Hastings

The Metropolis-Hastings algorithm provides a far more flexible framework for constructing transition kernels obeying the invariance condition eq. (5.16) than Gibbs sampling, and can be seen as an extension of the importance sampling

method eq. (5.7). Again the goal is to construct a transition kernel T which admits π as its invariant distribution and this is accomplished as follows: Starting from some point $x_t \in \mathcal{X}$, a new point $y \in \mathcal{X}$ (the proposal) is generated from a fixed axillary transition kernel Q which need *not* admit π as its invariant distribution. To correct for this discrepancy, y may be rejected (with some probability determined later) in which case x_{t+1} is set to x_t. Otherwise x_{t+1} is set to y. Surprisingly, the probability of accepting y turns out to have a simple analytical form. In full details the construction is as follows:

Assume the current state of the chain is $x_t \in \mathcal{X}$ and we wish to construct a transition kernel $T(x, \cdot)$ which determine the next state of the chain and admit π as the invariant distribution. In line with the above description, we consider a general transition kernel $Q(x, A)$, $x \in \mathcal{X}, A \subseteq \mathcal{X}$ which need *not* admit π as the invariant distribution. Assume for definiteness that Q takes the form

$$Q(x, dy) = q(y|x)\mu(dy) \tag{5.27}$$

where μ is the usual Borel measure and we have used the usual notation for conditional density for the function q of x and y to make the following more familiar. To avoid trivial complications, let $S = \{x : \pi(x) > 0\}$ be the support of π and assume for all $x \notin S$: $Q(x, S) = 1$. That is, when considering the Markov chain induced by Q, it always sends a point x which are not in the support of π into π's support. For simplicity, assume S contains more than a single point.

The *acceptance rate* $a(y; x)$ is now defined through

$$a(y; x) \equiv \begin{cases} \min\left\{\frac{q(x|y)\pi(y)}{q(y|x)\pi(x)}, 1\right\} & \text{if } q(y|x)\pi(x) > 0 \\ 1 & \text{if } q(y|x)\pi(x) = 0. \end{cases} \tag{5.28}$$

thus, the acceptance rate is a *function* of two variables. With these preliminary definitions we can proceed with the full method: Assume the current state of the chain is $x_t \in \mathcal{X}$ then x_{t+1} is generated by

- Generate $y \sim Q(x_t, \cdot)$
- Compute: $a(y|x)$
  - With probability $a(y|x)$ set: $x_{t+1} = y$
  - otherwise set: $x_{t+1} = x_t$.

It is instructive to write out the transition kernel induced by this procedure. First define the function t through

$$t(y; x) \equiv \begin{cases} q(y|x)a(y; x) & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases} \tag{5.29}$$

Then introducing

$$r(x) \equiv 1 - \int \mu(dy)\ t(y;x) \tag{5.30}$$

the full transition kernel can be written as

$$T(x, dy) \equiv t(y;x)\mu(dy) + r(x)\delta_x(dy). \tag{5.31}$$

By considering the different cases of eq. (5.28) and eq. (5.29) one obtains:

$$\pi(x)t(y;x) = \pi(y)t(x;y) \tag{5.32}$$

from which the balance condition eq. (5.16) follows from simple calculations [Tierney, 1994]

$$\begin{aligned}
\int \pi(dx)T(x,A) &= \int \pi(dx) \int_A T(x,dy) \\
&= \int \pi(dx) \int_A [t(y;x)\mu(dy) + r(x)\delta_x(dy)] \\
&= \left[\int \mu(dx) \int_A \mu(dy)\pi(x)t(y;x)\right] + \int_A \pi(dx)r(x) \\
&= \left[\int_A \mu(dy) \int \mu(dx)\pi(y)t(x;y)\right] + \int_A \pi(dx)r(x) \\
&= \left[\int_A \mu(dy)\pi(y)(1-r(y))\right] + \int_A \pi(dx)r(x) \\
&= \left[\int_A \pi(dx)(1-r(x))\right] + \int_A \pi(dx)r(x) \\
&= \int_A \pi(dx) = \pi(A)
\end{aligned} \tag{5.33}$$

The two terms in eq. (5.31) can be interpreted as the chance of moving from x to y in an accepted move plus the chance of having the move rejected and staying at x. As expected, convergence will depend on Q and more exactly on the relationship between Q and π, see [Nummelin, 1984, section 2.4] for more details.

The Metropolis-Hastings method depend on two separate innovations made 17 years apart. First by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [1953] for symmetric proposal kernels, that is, $q(x|y) = q(y|x)$ (notice the cancellation effect in eq. (5.28)) and later generalized to the present form by Hastings [1970]. The choice of acceptance function is not unique, however it is optimal in a number of circumstances [Peskun, 1973]. This leaves the choice of proposal kernel. Clearly setting $q(y|x) = \pi(y)$ is optimal in the sense of

providing i.i.d. samples from π and having acceptance rate 1. As with Gibbs sampling, a number of additional ideas and extensions are available. These include the *hit-and-run* algorithm [Smith, 1984], *multistage sampling* [Valleau and Card, 1972], diffusion-based methods for continuous problems such as the *Metropolis adjusted Langevin algorithm* [Roberts and Stramer, 2002], *Hamiltonian Monte Carlo* [Duane et al., 1987] and *Riemannian Monte Carlo* [Girolami and Calderhead, 2011], the *Multiple-try Monte Carlo method* [Liu et al., 2000], the *reversible-jump method* [Green, 1995] and a host of axillary variable techniques, hereunder the double metropolis-hastings algorithm for models where only an unbiased estimator of the normalization constant exist [Liang, 2010] and, for spin systems, the *Swendsen-Wang* algorithm [Swendsen and Wang, 1987]. Again this list is certainly not meant to be exhaustive and more details can be found in [Brooks et al., 2011].

The Metropolis-Hastings method is more generally applicable than Gibbs sampling, however constructing good proposal kernels is by no means an easy task. Specifically in the case of partition-based models most of the above techniques are not applicable because they are designed to treat different issues (such as growing parameter space or continuous problems), or because they are so general they still require novel ideas to be applicable to partition-based problems.

5.5 Adaptive Markov chain Monte Carlo

In most circumstances a MCMC proposal kernel Q (see eq. (5.27)) will depend on parameters which must either be set from a-priori information or by performing multiple runs for determining the optimal setting. An idea which is becoming more influential is to construct methods which attempts to learn these parameters online as samples $(x_t)_{t=1}^{\infty}$ from the Markov chain $(X_t)_{t=1}^{\infty}$ becomes available. The following treatment is based on Roberts and Rosenthal [2007]. Consider a family of transition kernels $\{T_\gamma(\cdot,\cdot)\}_{\gamma\in\mathcal{Y}}$ parameterized by a parameter γ in a space $\mathcal{Y}$ and assume all kernels have π as their stationary distribution in the usual sense of eq. (5.16). We will assume each kernel T_γ is irreducible and aperiodic, in other words, keeping γ fixed, T_γ acting upon an initial state x_0 will converge to π with probability 1.

The key idea in adaptive Markov chain Monte Carlo is to let γ at each iteration t depend on the past states of the chain and past state of γ. To this end, let Γ_t be a $\mathcal{Y}$-valued stochastic variable determining the value of γ at time t. The past information available at time t consists of the value of $(X_0,\ldots,X_t,\Gamma_0,\ldots,\Gamma_n)$.

More formally the available information $\mathcal{I}_t$ is the natural filtration

$$\mathcal{I}_t \equiv \sigma(X_1, \ldots, X_t, \Gamma_0, \ldots, \Gamma_t) \tag{5.34}$$

of the process $\{(X_t, \Gamma_t)\}_{t=0}^{\infty}$. Recall a filtration is a σ-algebra which contain all events that can happen up to time t, ie. all possible pasts. The joint state $\{(X_t, \Gamma_t)\}_{t=0}^{\infty}$ now depends on the past trough

$$\gamma_t \sim P(\Gamma_t \in \cdot \,| X_t = x_t, \mathcal{I}_{t-1} = I_{t-1}) \tag{5.35}$$

$$x_{t+1} \sim P(X_{t+1} \in \cdot \,| X_t = x_t, \Gamma_t = \gamma_t, \mathcal{I}_{t-1} = I_{t-1}) \equiv T_{\gamma_t}(x_t, \cdot\,). \tag{5.36}$$

It is the particular choice of eq. (5.35) that leads to a particular *adaptive Markov chain Monte Carlo* algorithm. The concept will be illustrated with a few simple special cases [Roberts and Rosenthal, 2007]. As a first example, if $\Gamma_t = \Gamma_0$ for all t the method reduce to ordinary Markov chain Monte Carlo. As a second example, if the choice of the current kernel Γ_t does not depend on the past values of $X_1, \ldots, X_t$ convergence is also guaranteed. As a third example, if there is some finite time τ after which the adaptation stops, i.e. $\Gamma_{\tau+t} = \Gamma_t$, the sampler also converges. The convergence of these schemes is perhaps not very surprising.

For a counter-example, consider a simple system consisting of two states, $\mathcal{X} = \{0, 1\}$ and two transition kernels. One transition kernel, the random kernel, choose a state at random, the other, the sticky kernel, stay in its current state with probability $1 > a > \frac{1}{2}$ and choose the other state with probability $(1 - a)$. Suppose we always choose the random kernel in state 0 and the sticky kernel in state 1. In this case clearly we will not sample the stationary distribution (which is uniform for both kernels) but the stationary distribution is now $\tilde{\pi}(0) = \frac{2-2a}{3-2a}$.

The above example is very degenerate, but it illustrates the transition kernel cannot deterministically depend on the current state. However if we impose the condition that the kernel change less and less between each iteration convergence can be guaranteed. This condition is specified in the following result [Roberts and Rosenthal, 2007]

Theorem 5.5.1. *Consider an adaptive Markov chain Monte Carlo algorithm $(X_t)_{t=1}^{\infty}$ on a space $\mathcal{X}$ with kernels T_γ for $\gamma \in \mathcal{Y}$ each with invariant distribution π. If the following conditions apply the algorithm is ergodic with stationary distribution π:*

Simultaneous uniform ergodicity: *For all $\varepsilon > 0$ there exist an N such that $\|T_\gamma^{(N)}(x, \cdot\,) - \pi(\cdot\,)\|_{var} \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.*

Diminishing adaptation: *Let the $\mathcal{I}_{t+1}$-measurable random variable D_t be defined as $D_t \equiv \sup_{x \in \mathcal{X}} \|T_{\Gamma_{t+1}}(x, \cdot\,) - T_{\Gamma_t}(x, \cdot\,)\|$. The diminishing adaptation property requires $\lim_{t\to\infty} D_t = 0$ (in probability).*

The first property is intended to avoid pathologies associated with infinite parameter spaces $\mathcal{Y}$. The second is the more important. Notice for instance it does not require the adaptation to stop or converge. To take the above example of the two-state system, and denoting the kernels T_0 and T_1, we may consider an adaptive method with transition kernel: $T_t = \cos(s_t)^2 T_0 + \sin(s_t)^2 T_1$, $s_t = \sum_{i=1}^{t} \frac{1}{i}$. Clearly $\lim_t s_t = \infty$, however from $|s_{t+1} - s_t| = (1+t)^{-1}$ the diminishing adaptation property follows by a simple analytical argument. The above result admits a number of important special cases and alternative formulation, most notably the simultaneous uniform ergodicity property can be relaxed [Roberts and Rosenthal, 2007, section 6], however the alternative conditions are more technical and it is an ongoing challenge to characterize the conditions under which adaptive methods converges.

The idea of a changing transition kernel has a long history [Gelfand and Sahu, 1994, Gilks et al., 1994]. An important contribution which closely resembled the above formulation of the problem as well as explicitly showed convergence to the correct stationary distribution was given by Haario et al. [2001] for a multivariate normal transition kernel. This quickly lead to many important generalizations and convergence results [Andrieu and Robert, 2001, Atchadé et al., 2005, Andrieu et al., 2006, Roberts and Rosenthal, 2007, Atchadé et al., 2010, 2009]. The past decade has seen an increase in application of adaptive Markov chain Monte Carlo. Noteworthy examples include *adaptive Metropolis-within-Gibbs* [Bai, 2009], *regional adaptation* and neighbour-based methods [Bai et al., 2011, Craiu et al., 2009], *adaptive Gibbs sampling* [Łatuszyński et al., 2013], the *Adaptive Equi-Energy Sampler* [Schreck et al., 2013] and *adaptive parallel tempering* [Araki and Ikeda, 2013] to mention some highlights from this growing litterature. See also Roberts and Rosenthal [2009], Liu [2008] for good overviews which covers many of these methods.

5.6 Remarks on convergence

A difficult problem when applying Markov chain Monte carlo methods is knowing how many iterations of the sampler are required before convergence results such as those in theorem 5.3.2 attains a particular precision. For a few problems (notably a hierarchical normal-means model) it is possible to obtain computational bounds [Rosenthal, 1995a,b], however these require sophisticated and very laborious analysis.

Setting the general question aside for a moment, we might consider the simpler question of how many iterations of the sampler is required before we can expect the current state of the sampler to be a random sample form the true posterior

distribution. Methods which guarantee samples from the true posterior are commonly denoted *exact sampling* techniques. This question too is very difficult to answer, however for certain problems and methods there exists general results using coalescent theory. In the simplest form, these methods revolve around starting chains in *all* possible states $\mathcal{X}$; not only does this require $\mathcal{X}$ to be finite, for a problem where exhaustive enumeration is feasible one could simply draw the samples explicitly. In a seminal contribution Propp and Wilson [1996] demonstrated this requirement can be softened to consider as few as two chains if the combination of transition kernel and $\mathcal{X}$ fulfilled certain properties and in the same work exact sampling was applied to a ferromagnetic Ising spin systems using Gibbs sampling and only two initial chains. These methods have subsequently been extended to certain continuous problems [Murdoch and Green, 1998], to more efficient samplers than Gibbs sampling on ferromagnetic Ising systems [Huber, 1999], other discrete systems [Del Genio et al., 2010, Carter et al., 2012] and as part of a sampler for double-stochastic systems [Murray et al., 2012], however their computational cost tend to be large and requires the system conform to strict requirements which has so far hindered widespread usage of exact sampling.

5.6.1 Assessing convergence in practice

If we consider a general realistic setting, we are often in the situation where the stochastic nature of the algorithms and our inability to tell which areas of the posterior distribution has a high value beforehand conspires to put us in the situation where it is not uncommon to have a method which not just take a *long* time to converge, it is not possible to tell how long time the sampler *should* run to converge and even if the sampler eventually begins to produce samples from the posterior distribution, we have no general way to tell this is the case. It is difficult to think of worse behavior for an algorithm which is guaranteed to converge.

It is useful to consider this problem from a practical perspective. Suppose we run a Markov chain for n iterations which gives the (coupled) states $(X_i)_{i=1}^n$. We can now choose $0 < m \leq n$ and an integer $s \geq 1$ and for some function f consider the stochastic variable

$$c_{n,s,m}[f] = \frac{1}{|S|} \sum_{i \in S} f(X_i), \qquad S = \left\{ s \left\lfloor \frac{i}{s} \right\rfloor : m \leq i \leq n \right\} \tag{5.37}$$

that is, the estimators based on samples from m to n with a spacing of s. m is known as the *burnin* time and is typically chosen as a fraction of n, for instance $n/2$. What we typically wish to know is how large to choose n, s and m such that

the above stochastic variable closely approximate the average based on $n_0 > 0$ i.i.d. samples from the invariant distribution π. In this case the Markov chain is said to *mix* or (with loosely applied terminology) to have *converged*. Posing the problem this way does not avoid any of the previously mentioned problems. What is typically done is to set up criterions to determine if the sampler has *not* run for sufficiently long time, under a common name these criteria are called convergence *diagnostics* or *statistics*. If for instance the values of $c_{n,s,m}[f]$ (for a suitable function f) appears to be correlated or if they change systematically (for instance if they tend to grow for the duration of the simulation) this is indicative that the sampler has not been evaluated for sufficiently long time or that the spacing s is to small.

Notice that by phrasing the problem as one of *deciding* the values of n, s and m and thereby putting up criteria which may or may not be triggered the problem is usually treated under decision theory or frequentistic statistics. Thus, convergence is usually treated as a null hypothesis and convergence statistics are commonly interpreted as *ruling out* convergence, *not* to provide a guarantee of convergence.

There exists a range of possible convergence diagnostics which can roughly be said to fall into two broad categories. The first are those which consider a single Markov chain evaluated for a very long time and consider the function values of a function f of X_i as a time series and use this quantity to assess convergence. A plot of such a time series is known as a *trace plot*.

If f is a real function, we may expect the mean of $f(X_i)$ to agree between the first and last third of the time series. Assuming normality, this is a univariate test which can be assessed using Z-score to produce a simple convergence check [Geweke et al., 1991]. Related to this idea is the computation of autocorrelation time which can subsequently be used to estimate the effective sample size [Kass et al., 1998].

A potential issue with diagnostics that only consider a single chains may be illustrated with the following example. Consider two chains with the same stationary distribution and the following loose description of the behavior of the chains: We assume the stationary distribution has two peaks in $\mathcal{X}$ such that the chains tend to get stuck at each peak. Assume the first chain mix very poorly and will invariantly get stuck at one peak and change peaks at a frequency orders of magnitude lower than the longest realistic simulation time. The other chain is slightly better and, while it still get stuck at each peak for a long time, it will typically skip between the peaks a few times during each simulation. While the first chain is no doubt worse than the second from any objective standpoint, a comparison of the two chains by a single-chain statistic influenced by the value at the two peaks will suffer from a Dunning-Kruger

effect [Kruger and Dunning, 1999] in that the convergence diagnostic indicate the first chain mix far better than the second *because* it mix so poorly it is unable to uncover how poorly it mixes.

The above problem can be somewhat alleviated by running multiple chains in parallel and investigating the behavior of the time series induced by f *across* different chains. This is the idea behind the Gelman-Rubin *reduction coefficient* which compares the variance *within* each chain to that which is computed *between* the entire pool of chains. This is collected in a single number, the reduction coefficient, which (assuming the chains mix) can be expected to converge to 1 [Gelman and Rubin, 1992]. Clearly this method depends on the different chains being initialized in states which are very spread out, i.e. such that if the configuration space $\mathcal{X}$ consist of isolated peaks one has a good chance of finding them as local minima. In addition to this difficulty all the preceding issues regarding the choice of f applies.

Many additional methods for assessing convergence exist and the reader is referred to the reviews Cowles and Carlin [1996], Brooks and Gelman [1998] and Plummer et al. [2006] for further references and discussion. In conclusion, there exist no general-purpose test for convergence or divergence of Markov chains. Within existing methods, between-chain diagnostics such as Gelman-Rubin statistics seems to provide the better starting point, especially when comparing methods where some may mix very poorly. From a practical perspective, especially when developing sampling methods, visual inspection of the trace plots often offers more insight in the behavior of the chain than a convergence statistics, with the obvious drawback of not offering a quantitative comparison.

Additionally, suppose one considers a practical application where two Markov chains is applied to the same problem. Suppose it is the case neither chain converges, however one chain may performs far worse than the other. In this situation one would tend to consider smaller problems, however it may be the difference (in terms of proposing new states) between the methods is not very great on smaller problems (say one method manipulates several variables jointly, then this will presumably be less important when there are fewer variables to manipulate), or that their time complexity is different and finally, the very pragmatic considerations other researchers will, damn the torpedoes, disregard convergence and apply the method to their particular problem and it is in *that* setting they are really interested in the performance. However in this large-system limit it may be all convergence diagnostics will (and should) agree neither chain has converged and trace plots may be easier to interpret.

A general feature of convergence diagnostics is the reliance on a real functions f. While this is an excellent starting point if the problem consists of a fixed set

of continuous variables where a few can be expected to be very hard to sample, for discrete structures such as trees, partitions or graphs any such single number may be highly deceptive. To take partitions, one could consider the indicator variable $\delta_{z_i - z_j}$ of the assignment of two observations i and j. If the problem contains latent structure (which is hopefully the case!) most of these indicator variables may be expected to be constant, and even when this is not the case, it is typically the joint assignment of a subset of the variables which is crucial for exploration. Alternatively, one could consider the number of blocks in a partition as f; however in our experience this statistics too can be highly misleading because small, spurious, blocks and large important block contribute equally to the statistic. Similar problems may be raised for other univariate statistics and good literature on this problem has not been found (we discuss this briefly in Herlau et al. [2014a]); mind we do not consider our contribution to have made progress on this problem.

5.7 Sampling partitions

In this section, we consider a specialized application Markov chain Monte Carlo the problem of sampling models based on partitions. The aim is to motivate the adaptive sampling proposal, *Adaptive Reconfiguration Moves for Efficient Markov Chain Sampling* [Herlau et al., 2014a], however the chapter will focus on several related ideas that nevertheless did not work; we believe these other failed attempts are worthwhile to mention despite being unsuited for publication. While we will introduce much of the notational machinery required to describe our method we will not describe it in much depth since these details can be found in Herlau et al. [2014a].

5.7.1 Basic notation

The notation we will use will be consistent with what has previously been introduced, but with some minor extensions. We will continue to use π to denote a partition of a set $\mathcal{X}$ of n elements. Typically we will choose $\mathcal{X} = [n] = \{1, 2, \ldots, n\}$. Let $K = |\pi|$ denote the number of blocks in the partition and recall the notation

$$\pi = \{B_1, B_2, \ldots, B_K\} \quad \text{where} \quad B_k \subseteq \mathcal{X}. \tag{5.38}$$

A partition π of $\mathcal{X}$ induces an equivalence relation $[\cdot]_\pi$ and we will write:

$$[i]_\pi = \{j : i, j \in B_k \text{ for some } k\} \tag{5.39}$$

for the block of the partition π containing i. By a simple (or conjugate) partition based model we mean a model where any block-specific parameters can be integrated out. Letting Y denote the data we may write the density of such a model as

$$p(Y, \pi). \tag{5.40}$$

In the following the function $p(Y, \pi)$ is abbreviated by $q(\pi)$.

5.7.2 The infinite relational model

The canonical example of such a model is the *infinite relational model* (IRM) for network data Kemp et al. [2006], Xu et al. [2006]. The IRM is trivially obtained form the stochastic block model eq. (2.90) of chapter 2 by replacing the categorical distribution on π with the Chinese restaurant process eq. (4.41); for completeness the full generative model becomes:

$$\begin{aligned}
& \pi && \sim \mathrm{CRP}([n], \alpha) && && \text{(5.41a)}\\
\text{for } 1 \leq \ell \leq k \leq |\pi| \quad & \theta_{\ell k}|z && \sim \mathrm{Beta}\,(b_1, b_2) && && \text{(5.41b)}\\
\text{for } 1 \leq i < j \leq n \quad & A_{ij}|\theta, z && \sim \mathrm{Bernoulli}\,(\theta_{\ell k}) && \text{st. } i \in B_\ell, j \in B_k. && \text{(5.41c)}
\end{aligned}$$

Collecting the terms and integrating allows us to express the joint likelihood to be sampled, $q(\pi)$:

$$p(A, \pi) = \int d\theta \; p(\pi|\alpha)p(\theta|\pi)p(A|\theta, \pi) \tag{5.42}$$

$$= \prod_{1 \leq \ell \leq k \leq |\pi|} \frac{B(b_1 + N^+_{\ell k}, b_2 + N^-_{\ell k})}{B(b_1, b_2)} \frac{\Gamma(\alpha)\alpha^{|\pi|}}{\Gamma(n + \alpha)} \prod_{b \in \pi} \Gamma(|b|) \tag{5.43}$$

$$\equiv q(z) \tag{5.44}$$

where $B(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the beta function and

$$N^+_{\ell k} = \sum_{i<j} A_{ij} 1_{B_\ell}(i) 1_{B_k}(j), \qquad N^-_{\ell k} = \sum_{i<j} (1 - A_{ij}) 1_{B_\ell}(i) 1_{B_k}(j) \tag{5.45}$$

are the pseudo counts.

5.7.3 Operations on partitions

We will again make use of common operations on the partitions. Firstly, all observations in the blocks of a partition π is written as $\cup\pi \equiv \cup_{b \in \pi} b$ and for a

block A, recall that $\text{proj}_A \pi$ denotes the partition obtained by restricting each block to A and removing any blocks which do not overlap with A:

$$\text{proj}_A \pi = \{B \cap A : B \in \pi \text{ and } B \cap A \neq \emptyset\} \tag{5.46}$$

Assume $I = \{I_1, \ldots, I_{|I|}\}$ is a sequence of non-overlapping subsets of $\mathbb{N}$ such that $I \subseteq \pi \cup \{\emptyset\}$ and suppose $A \subset \mathbb{N}$, $A \neq \emptyset$ and either (i) there exists a k such that $A \subseteq B_k$ or (ii) A is disjoint from the set of all observations in π, $A \cap (\cup\pi) = \emptyset$. We then define a (restricted) Gibbs sweep on π as the probability kernel

$$\kappa_{I,A}(\pi'|\pi) \tag{5.47}$$

in which the set of observations A is assigned to each block of π contained in I and, assuming $\emptyset \in I$, a new block only containing A. These partitions are then sampled with probability proportional to q in a Gibbs sweep.

To be more specific, first consider the set of partitions $(\pi^{(1)}, \ldots, \pi^{(|I|)})$ obtained from π by removing the elements in A from π and then adding A to each block in I. Specifically for $k = 1, \ldots, |I|$ set

$$\pi^{(k)} = \{I_k \cup A\} \cup \text{proj}_{\mathcal{X} \setminus (I_k \cup A)} \pi, \quad \mathcal{X} = \cup\pi \quad \text{and} \quad a^{(k)} = q(\pi^{(k)}) \tag{5.48}$$

and then setting $\pi' = \pi^{(k)}$ where k is sampled from a multinomial distribution with weights

$$\left(\frac{a^{(1)}}{\sum_{k=1}^{|I|} a^{(k)}}, \ \ldots, \ \frac{a^{(|I|)}}{\sum_{k=1}^{|I|} a^{(k)}} \right). \tag{5.49}$$

To recover the standard Gibbs sweeps considered in the previous section I is set equal to all available blocks in π and a new block (corresponding to the empty set). To this end we define

$$\kappa_A(\pi'|\pi) \equiv \kappa_{\pi \cup \{\emptyset\}, A}(\pi'|\pi) \tag{5.50}$$

and the standard Gibbs kernel for an observation $i \in \mathcal{X}$ can be written as

$$\kappa_{\{i\}}(\pi'|\pi). \tag{5.51}$$

5.7.4 Split-merge sampling

The incremental nature of Gibbs sampling makes it prone to get stuck in local modes where it either over or under estimates the true number of blocks Celeux et al. [2000]. For this reason it is natural to consider alternative methods based

on Metropolis-Hastings transitions which allow one to make larger changes in the space of partitions in one operation.

One particular set of moves are *split-merge* moves where either a single block is split into two new blocks or two blocks is merged into a single block *and in both cases all other blocks are kept fixed.* While the merge step is unique, there are multiple ways to perform the split step. One of the most popular is the split-merge method of Jain and Neal [2004]. The method proposes a split configuration by randomly selecting two observations i, j and then randomly distributing the observations assigned to the block(s) containing i, j between two new blocks, then perform a number of Gibbs updates restricted to only moving observations between these two new blocks to obtain near equilibrium split configuration and a final Gibbs update to arrive at a split-proposal. The key observation, which may appear quite surprising, is one *only* needs to consider the *last* restricted Gibbs sweep when computing the corresponding acceptance probability eq. (5.28).

To introduce some relevant notation we might consider the full transition kernel as follows: First, the kernel $T(\pi^*|\pi)$ (for definiteness assume $[i]_\pi = [j]_\pi$) is a mixture:

$$T(\pi^*|\pi) \equiv p(\phi)T_\phi(\pi^*|\pi) \tag{5.52}$$

for an index set ϕ. Clearly if each kernel T_ϕ satisfy the balance condition eq. (5.16) so will T; and if just one T_ϕ (assuming $p(\phi) > 0$ and ϕ is discrete) is ergodic so is the full kernel. Now, the transition kernel is decomposed into a proposal step $t_\phi(\pi|\pi^*)$ and an acceptance step $a_\phi(\pi^*|\pi)$ as in eq. (5.28). First, the background information ϕ (in the simplest form) consist of two (distinct) vertices i, j and a random initial partition of S into two blocks such that i and j belong to different blocks. Denote by π^l the *entire* path of launch state, that is, the initialization of i, j into two different blocks (obtained deterministically by restricting the partition of S in ϕ to the vertices in $[i]_\pi \cup [j]_\pi$) followed by the application of a number of restricted Gibbs transition kernels to obtain the final launch state. Assuming $t^l_\phi(\pi^l|\pi)$ is the density of this procedure and let $t_\phi(\pi^*|\pi^l, \pi)$ denote the kernel corresponding to the application of a *single* restricted Gibbs kernel to obtain the final state π^* from the (last) state of the path π^l the acceptance probability is then defined as:

$$a_\phi(\pi^*|\pi^l, \pi) = \min\left\{1, \Delta\mathcal{L}\right\}, \quad \Delta\mathcal{L} = \begin{cases} \frac{q(\pi^*)}{q(\pi)t_\phi(\pi^*|\pi^l,\pi)} & \text{if } [i]_\pi = [j]_\pi \\ \frac{q(\pi^*)t_\phi(\pi|\pi^l,\pi^*)}{q(\pi)} & \text{if } [i]_\pi \neq [j]_\pi. \end{cases} \tag{5.53}$$

Compared to the definition of the Metropolis Hastings procedure in eq. (5.28) this operation differs in that the probability of the launch state is missing from

the acceptance probability. In this notation, for $\pi^* \neq \pi$, we may express the transition kernel by marginalizing over all intermediate paths π^l:

$$T_\phi(\pi^*|\pi) = \int d\pi^l \; t^l_\phi(\pi^l|\pi) t_\phi(\pi^*|\pi, \pi^l) a_\phi(\pi^*|\pi^l, \pi) \tag{5.54}$$

with the caveat $t_\phi(\pi^*|\pi, \pi^l)$ deterministically merge the two blocks $[i]_\pi$ and $[j]_\pi$ if $[i]_\pi \neq [j]_\pi$.

To show convergence, consider the case $\pi^* \neq \pi$. It suffices to show the detailed balance condition for each ϕ:

$$q(\pi) T_\phi(\pi^*|\pi) = q(\pi^*) T_\phi(\pi|\pi^*) \tag{5.55}$$

which can easily be seen to imply the balance condition eq. (5.16) by summing over π^* and inserting into eq. (5.52). However by simple insertion of eq. (5.54) into the right and left-hand side of the detailed balance condition eq. (5.55) it is easily seen to be (assuming $[i]_\pi = [j]_\pi$ with the other case following by symmetry):

$$\begin{aligned}
&\int d\pi^l \; q(\pi) t^l_\phi(\pi^l|\pi) t_\phi(\pi^*|\pi^l, \pi) a_\phi(\pi^*|\pi^l, \pi) \\
&= \int d\pi^l \; q(\pi) t^l_\phi(\pi^l|\pi) t_\phi(\pi^*|\pi^l, \pi) \min\left\{1, \frac{q(\pi^*)}{q(\pi) t_\phi(\pi^*|\pi, \pi^l)}\right\} \\
&= \int d\pi^l \; \min\left\{q(\pi) t^l_\phi(\pi^l|\pi) t_\phi(\pi^*|\pi, \pi^l), q(\pi^*) t^l_\phi(\pi^l|\pi)\right\}.
\end{aligned} \tag{5.56}$$

And the reverse rate:

$$\begin{aligned}
&\int d\pi^{l'} \; q(\pi^*) t^{l'}_\phi(\pi^{l'}|\pi^*) t(\pi|\pi^{l'}, \pi^*) a_\phi(\pi|\pi^{l'}, \pi^*) \\
&= \int d\pi^{l'} \; q(\pi^*) t^{l'}_\phi(\pi^{l'}|\pi) t(\pi^*|\pi^{l'}, \pi) \min\left\{1, \frac{q(\pi) t_\phi(\pi^*|\pi^{l'}, \pi)}{q(\pi^*)}\right\} \\
&= \int d\pi^{l'} \; \min\left\{q(\pi) t^{l'}_\phi(\pi^{l'}|\pi^*) t_\phi(\pi^*|\pi^{l'}, \pi), q(\pi^*) t^{l'}_\phi(\pi^{l'}|\pi^*)\right\}
\end{aligned} \tag{5.57}$$

The crucial property is the information ϕ will, by construction, erase the difference of π and π^* as far as the transition kernel is concerned, ie. such that $t^l_\phi(\pi^l|\pi^*) = t^l_\phi(\pi^l|\pi)$. Detailed balance now follows by simple relabelling.

5.8 Other methods for sampling partitions

To understand the proposed extensions it is important to understand the limitation of the Gibbs sampling. A view which is intuitively tempting but misleading is as follows: while each assignment of π may change with relatively low probability, that probability must be positive and so, given sufficiently long time, changes in the coordinate must accumulate and drive the state of the Markov chain to different areas of state space.

In practice, for the IRM model even on small problems, the following very crude view seem to better reflect what happens burnin: 95% of the observations in π are fixed and remain fixed for the duration of the chain. The other 5% change assignments during the simulation and tend to do so quite often, however their change in assignment consist in jumping between the same fixed block structure as defined by the remaining 95% of the observations. Different restarts of the simulation will find different such patterns. As a consequence, there is no exploration of state space.

Evidently, the problem has to do with changing the assignment of *blocks* of observations rather than single observations in one step. This is naturally not a new observation [Celeux et al., 2000, Jain and Neal, 2004], however the degenerate behavior of Gibbs sampling on the IRM model is so striking even quite naïve samplers which *do* move blocks of observations may have an advantage over simple Gibbs sampling.

Idea 1 The idea of the first method is quite basic: Given a partition $\pi = \{B_1, \ldots, B_K\}$, randomly select a block B_k and a random subset $b \subset B_k$ and re-assign the observations in the subset b using a Gibbs sweep

$$\pi^* \leftarrow \kappa_b(\cdot|\pi) \tag{5.58}$$

This superficial description leaves open how to appropriately select b; one way is the following scheme based on coagulation theory. First, for a partition $\pi = \{B_1, \ldots, B_K\}$ and a second partition c of $[K]$ we define the *coagulation* as the partition π':

$$\pi' = \mathrm{Coag}_c(\pi) \equiv \left\{ \bigcup_{k \in S} B_k : S \in c \right\} \tag{5.59}$$

and we say that π *is coagulated by* c.

Secondly, we define the CRP-Coag$_{(\alpha,d)}$ operation as a Markov transition kernel on the set of partitions indexed by two parameters $0 \leq d < 1$ and $\alpha > 0$. Acting

on a partition π, it's action CRP-Coag$_{(\alpha,d)}(\Pi = \pi', \pi)$ is defined through the two-step procedure using the two-parameter CRP of eq. (4.44):

$$c \sim \text{CRP}\big(\big[|\pi|\big], \alpha, d\big) \tag{5.60a}$$

$$\pi' = \text{Coag}_c(\pi). \tag{5.60b}$$

It now holds that if π is distributed as a CRP$([n], d, \alpha)$ and π' conditional on π as a CRP-Coag$_{(\alpha/d,0)}(\cdot, \pi)$ then π' is distributed as a CRP$([n], 0, \alpha)$. We write the joint density of the two partitions obtained by these two procedures as $p_{\text{Coag}}(\pi, \pi'|d, \alpha)$. This is a very partial statement of more general results on Markov chains on the space of partitions known as coagulation-fragmentation processes. See for instance the comprehensive references Pitman and Picard [2006, chapter 5],Pitman [1999, 2002], Bertoin [2006], Ho et al. [2006], James [2009] for general treatments as well as Elliott and Teh [2012], Buntine and Hutter [2010] for particular statements of these results in the discrete setting.

Returning to the problem of creating a sampler, assume $q(\pi') = p(Y, \pi') = p(Y|\pi')p(\pi')$ is the partition-based problem. If we assume $p(\pi')$ is distributed as CRP$([n], 0, \alpha)$ we may consider

$$p(Y, \pi, \pi') = p(Y|\pi')p_{\text{Coag}}(\pi, \pi'|d, \alpha) \tag{5.61}$$

for any $0 < \alpha < 1$ it follows $q(\pi') = \int d\pi \; p(Y, \pi, \pi')$ by previous discussion and so we may consider a sampling method where the blocks in the *refined* partition π are moved using Gibbs operations.

It turn out this method is very competitive compared to split-merge sampling for the few problems where the method was applied. This is surprising since the method is plainly absurd: The refinement of each block of π' given by π is *not* informed by the likelihood and will be suboptimal in nearly all cases. The only reason why this method works is in a few cases the refinement will, by chance, allow energetically favorable moves and this class of moves also encompass the situation where a subset of observations changes *between* two blocks.

Idea 2 Given the observation in the past section it is natural to consider various ways to inform the subpartition of the actual likelihood. To this end we define the following: For two partitions π^a and π^b define the coarsest common refinement as

$$\pi' = \pi^a \vee \pi^b \equiv \left\{ A \cap B : A \in \pi^a, B \in \pi^b, A \cap B \neq \emptyset \right\}. \tag{5.62}$$

We may now consider the following sampling procedure: Suppose we have a set of Markov chains for $q(\pi)$ (or more formally, a single Markov chain on the product space). If we consider the current state of two chains, π^a and π^b and

compute $\pi' = \pi^a \vee \pi^b$ we may for any i (for instance chosen at random) consider the Gibbs sweep where the set of observations $[i]_{\pi'}$ is jointly assigned in (for instance) π^a. This procedure does *not* define a valid Markov chain since when computing the reverse probability $[i]_{\pi'}$ may have changed. However if one ignore this problem the resulting method is much more powerful than the simple random method (Idea 1).

Idea 3 Since the above method seems very efficient we explored some ways to make the construction more principal. One attempt is a tempered sampling approach. Suppose some of the chains are tempered by the transformation: $q^{(\beta)}(\pi) \propto \exp(\beta \log q(\pi))$. Suppose 3 chains are selected, π^a, π^b and π^c such that π^a and π^b are from the untempered distributions while π^c is from the tempered distribution with eg. $\beta = \frac{1}{2}$. If we now consider $\pi' = \pi^a \vee \pi^b \vee \pi^c$ and Gibbs sample $[i]_{\pi'}$ in *all* three distributions from their joint distribution (notice this can be computed by considering the corresponding three Gibbs updates independently) under the restriction that for the new states, $\pi^{a*}, \pi^{b*}, \pi^{c*}$ and π'^* we have $[i]_{\pi'} = [i]_{\pi'^*}$ then this define a valid proposal operation on the joint space. The method can trivially be extended by considering several values of β until β is so low the problem is easily sampled with Gibbs sampling and this was the method we considered.

The idea behind this proposal is innovations at the hard-to-sample $\beta = 1$ temperature can be accepted by letting the higher-temperature chains incur the potential cost (in terms of likelihood) of a sub-optimal assignment of the block $[i]_{\pi'}$ while the benefit (again in terms of likelihood) may be enjoyed by the low-temperature chains. In addition one can hope innovations happening at the easier-to-sample chains of higher temperature, being from a *different* but *related* problem, might allow better exploration.

While this is an attractive idea, the simple fact is it does not appear to work very well. When the cost of the many tempered chains is taken into account (2 to 4 temperatures at temperature intervals of $\beta = 0.5^k$) seemed to be optimal) the method was only slightly better than Idea 1, the random two-stage coagulation moves and worse than Idea 2. It should be mentioned on the considered problems it appeared to performed better than Split-Merge sampling.

5.8.0.1 Discussion

A clear limitation of the preceding methods, including split-merge sampling of Jain and Neal [2004], is the reliance on the reassignment of a single block of variables. Consider the ideal case of a split-merge operation: There exist two

partitions π^a and π^b with $q(\pi^b) \geq q(\pi^a)$ such that they only differ in that a block in π^a corresponds to two blocks in π^b:

$$\pi^a = \{A_1' \cup A_1''\} \cup \{A_2, \ldots, A_K\}, \quad \pi^b = \{A_1'\} \cup \{A_1''\} \cup \{A_2, \ldots, A_K\}. \quad (5.63)$$

By comparing the output of many chains it was observed it is *very often* the case there are some small but crucial difference between π^a and π^b. Say instead of simply splitting the block $A_1' \cup A_1''$ in π^a into A_1' and A_1'' in π^b, a few observations from $A_1' \cup A_1''$ had to enter a third block in π_b while it may be the case A_1' also contained a few observations from a fourth block. While these differences most often involved very few observations, experiments indicated it was sufficient to guarantee there could often be *no* single split into a partition π^b which were favorable in the sense $q(\pi^b) \geq q(\pi^a)$. Since Gibbs sampling very often perform no appreciable exploration, this appeared to be a very important contribution to the poor mixing of split-merge sampling.

A second important factor is the requirement the number of components must increase or decrease by 1 in split-merge sampling. While the block count do very nearly always fluctuate, the *larger* reconfigurations are very often of the form where one set of observations move from one larger block to another and for many simulations it appeared no intermediate split or merged configuration could be found, effectively guaranteeing these moves never occurred.

In addition to these serious problems split-merge sampling has two features that almost guarantees a low acceptance rate. The simplest is any observations selected at random should nearly never be split or merged (compared to their current assignment) and the second is the acceptance rate eq. (5.53) can be surprisingly low even when "correct" observations are selected. To see this, consider the π^a and π^b example and consider the case where, by chance, two observations in B_1 and B_2 has been selected and a launch state π^l (recall this corresponds to a split partition) has been obtained. The accept rate then contains the term $t_\phi(\pi|\pi^l)$, a product of restricted Gibbs kernels, however there may be strong correlation amongst the assignment of observations to the two partitions in π^l making the density (and thus accept rate) very low.

5.8.1 Adaptive reconfiguration moves

The primary idea behind our proposal in Herlau et al. [2014a] is to enlarge the space of possible transitions to include more than simple split and merge operations. To put it in a deflationary way, this is done by dressing up a method *simpler* than the split merge method of Jain and Neal [2004]. We will first briefly review of the construction in a simplified form and then provide the full

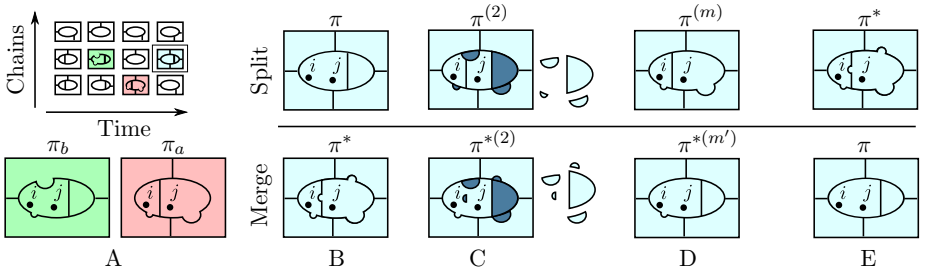

Figure 5.1: A single reconfiguration move applied to the partition π and observations i, j shown in (B). First, assume partitions π_a, π_b are chosen from the past states of $S = 3$ chains (A). The coarsest common refinement of π, π_a and π_b is used to construct the initial split, and the subblocks of the blocks containing i, j in any of the partitions are removed (C). The removed subblocks are Gibbs sampled into the partition giving partition $\pi^{(m)}, m = 6$ (D). Finally all observations (subject to restrictions discussed later) are allowed to move either from outside the blocks containing i, j and into the blocks containing i, j, or from inside the blocks containing i, j and outside creating the final partition π^* in (E). The bottom row shows the same process applied to create the reverse (merge) operation.

construction. An illustration of the final construction, *adaptive reconfiguration moves* (ARM) can be found in fig. 5.1 (figure and caption taken from Herlau et al. [2014a]) The method assumes we are given partitions π_a and π_b and two observations i, j such that $[i]_{\pi_a} \neq [j]_{\pi_a}$ and $[i]_{\pi_b} = [j]_{\pi_b}$ in addition to the current partition π. Suppose $[i]_\pi = [j]_\pi$ for specificity. A new partition π^* is then constructed by first computing the coarsest common refinement of π, π_a and π_b, $\pi' = \pi \vee \pi_a \vee \pi_b$, removing observations in $[i]_\pi \cup [j]_\pi$ and using the blocks $[i]_{\pi'}, [j]_{\pi'}$ as the initial split.

The removed subblocks (the subblocks in π' which have been removed) are then Gibbs sampled into the partition and finally all other observations are allowed to move either from outside the blocks containing i, j and into the blocks containing i, j, or from inside the blocks containing i, j and outside creating the final partition π^*.

The partitions π_a and π_b are selected at random from the past states of the current and other chain. As this set grows it is in this sense the method become an example of adaptive Markov chain Monte Carlo.

While the set of partitions which can be reached is not made larger by moving

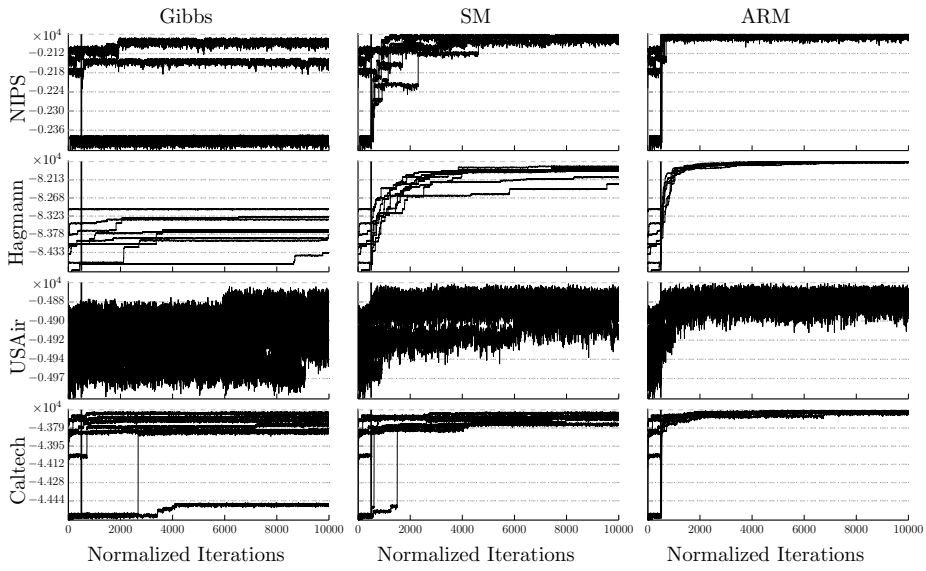

Figure 5.2: Trace plots of log likelihood for Gibbs (G), split-merge (SM), and ARM sampling for four network datasets. All simulations are based on running $S = 8$ chains using Gibbs sampling for 500 iterations, then continuing with Gibbs, SM or ARM sampling for up to 10000 iterations. The x-scale is normalized and represents the same computational effort.

subblocks initially (step (C)) rather than single observations, the use of subblocks allows faster thermalization giving rise to higher acceptance rate. As shown in Herlau et al. [2014a], the acceptance rate may be up to an order of magnitude larger than split-merge sampling.

Naturally this brief discussion omits several details necessary to get detailed balance. These mainly revolving around being able to compute the transition density from one state to another, $t(\pi^*|\pi)$, in the accept rate eq. (5.28) which impose a very slight restriction to the move class. The reader is referred to Herlau et al. [2014a] for these details as well as a proof of convergence.

In terms of performance the method has quite striking benefits compared to split-merge sampling at least for the considered models and datasets, see fig. 5.2 for a reproduction of the trace plot for the IRM model for selected network datasets.

Chapter 6

Networks

A *network* is a term with no single agreed upon definition. A network typically denotes a collection of units which are interacting or otherwise related in a quantifiable manner. Canonical examples of networks include the internet [Faloutsos et al., 1999, Albert et al., 1999], where the network consists of computers linked either by their interaction or physical connectivity, a society where humans interact either by physical proximity or social acquaintance [Eagle et al., 2009, Eagle and Pentland, 2006, Onnela et al., 2007, Gonzalez et al., 2008, Moody, 2004] or biology where for instance chemical components in a cell interact through relationships such as binding or catalysis [Barabasi and Oltvai, 2004, Combet et al., 2000, Jeong et al., 2001]. While examples of networks are often mapped onto simple graphs, all common extensions to simple graph may be considered such as hypergraphs, weighted networks, temporal changes and processes which occur on or otherwise involve graphs. Importantly, sometimes the word network is used to refer to something quite different from a simple graph, for instance a brain network sometimes denote spatially overlapping blobs of brain tissue [Bullmore and Sporns, 2009].

6.1 Subjects of network science

The study of networks may roughly be divided into the following six areas loosely adapted from the five areas of Newman [2010].

Zoology of networks Networks are interesting because many systems are believed to be well described as networks. Accordingly, networks are studied empirically where the goal is to map physical systems to networks and then describe and classify these networks based on their properties. The references provided in the introductory section are examples of this type of work.

Combinatorics and mathematical studies Networks was historically first studied as combinatorial objects. Most definitions used in network science was motivated within this are (graphs, trees, degree, cycle, etc.) and it contain a large set of important theoretical results in combinatorics such as conditions under which a coloring is possible (of edges or vertices), counting of unique trees or graphs and a range of other problems [Harris et al., 2008, Townsend, 1987, Chartrand et al., 2010].

Growth and physics Nearly everything that has to do with networks has been invented or re-invented within the complex physics community. A novel approach to the study of networks we associate with physics is the mindset of statistical physics, namely how the microstructure define macrostructure in networks and, as a natural concern, what the relevant microstructure and macrostructure *is*. In terms of macrostructure it has typically been properties such as degree-distribution, average paths lengths, reciprocity, modularity, triangle/motif statistics and so on that has played a central role with a large focus on when these properties follow a particular distribution such as a scale-free distribution. Microstructure has sometimes consisted of (a subset of) the above properties, but also the focus on growth mechanisms of networks. The research question commonly posed is: Can certain large-scale structures be identified as common in networks and can these in turn can be explained by simpler assumptions on the micro-structure [Erdös and Rényi, 1959, Newman, 2010, Barabási and Albert, 1999, Girvan and Newman, 2002, Albert and Barabási, 2002].

Networks Algorithms This study, which we loosely identify with computer science, is concerned with algorithms that take a network as an input and compute certain properties of the network, for instance identifying all subgraphs of a certain type or find the minimal cut-set. Concerns are thus network representation, runtime, convergence and so on [Bang-Jensen and Gutin, 2007, Jungnickel and Schade, 2005],

Statistical Modelling This label covers typical statistical tasks such as predicting missing data or otherwise infer parameters in statistical models. This will be the subject of this chapter.

Processes on networks Whenever one feel a particular task involving a single network becomes too simple, one can consider processes taking place on or involving a network. Examples include percolation theory for networks [Callaway et al., 2000], spreading phenomena such as information or epidemics [Pastor-Satorras and Vespignani, 2001, Newman, 2002], search and routing on a network [Stoica et al., 2001] and also other applications involving networks such as games [Nisan, 2007] or statistical models [Koller and Friedman, 2009]. Needless to say networks are interesting insofar as they *in some way* relate to a problem in this category.

While we have tried to break the above discussion into elements which are somewhat divided between interests and fields, it should be emphasized many of the above references are interdisciplinary. This section will be limited to a single of the above topics, statistical modelling of networks, and within this topic we will in the main only consider exchangeable models of random arrays where the latent structure is easily expressed using the Aldous-Hoover representation theorem 4.3.1 of chapter 4. A reader interested in a broader picture of network modelling may consult *"Networks"* [Newman, 2010] which to our knowledge is the most comprehensive reference; other noteworthy tutorials from various perspectives include the works of [Orbanz and Roy, 2013, Newman, 2003, Durrett, 2007, Goldenberg et al., 2010, Fienberg, 2012] and Schmidt et al. [2012].

An important topic not covered in this section is validation of network models. The most popular validation method is link prediction, typically as measured by AUC score or predictive likelihood [Schmidt et al., 2012], however the details often differ between publications and we will not attempt to cover the various approaches here.

6.2 Bayesian modelling of networks

Models for networks can roughly be divided into three categories. Those which are probabilistic but not exchangeable, those which are probabilistic and exchangeable and those which are neither. By exchangeable we fully restrict ourselves to exchangeable matrices and the Aldous-Hoover representor theorem 4.3.1 [Orbanz and Roy, 2013]. To reiterate, the Aldous-Hoover theorem states a random simple graph (ie. binary, symmetric and without self-edges) (X_{ij}) is jointly exchangeable if and only if there exists a random function

$W : [0,1]^2 \mapsto [0,1]$ (with zero diagonal) such that the presence or absence of a link x_{ij} is sampled from

$$x_{ij}|W,(U_i) \sim \text{Bernoulli}(W(U_i,U_j)) \tag{6.1}$$

for i.i.d. random variables $U_i \sim \text{Uniform}([0,1])$.

One important intuition to take away from the Aldous-Hoover theorem is it shows an exchangeable array can be decomposed into three parts: Firstly, the observational distribution. In eq. (6.1) this was the Bernoulli distribution. Secondly, the structural assumption that the parameter of the observational distribution can be decomposed as (random) vertex-specific properties (in eq. (6.1) this is the set of scalars (U_i), $U_i \in [0,1]$) and thirdly, a (random) structure which indicates how the vertex-specific properties interacts (in eq. (6.1) this was the random function W). Together these three parts defines the structural assumption for a particular exchangeable model of random simple graphs. Accordingly, if we define the (random) parameter(s) for edge (ij) as:

$$W_{ij} \equiv W(U_i,U_j). \tag{6.2}$$

Any particular model can be described by indicating the observational distribution and the structural assumptions. We will loosely call this an *Aldous-Hoover type* representation and the following sections will discuss a number of models by specifying the parametrization in eq. (6.2).

This is not to say all there is to say about a particular model is said by specifying the representation of W_{ij} and the discussion below will often omit important details. Otherwise all models considered would be exchangeable and this is not the case. It is nevertheless an intuitive way to get an overview of the structural assumption of the models and their relationship.

The particular parameterizations of W_{ij}, or models, can again be roughly divided according to the choice of space for the latent vertex-specific parameters (U_i) which, in conjunction with the choice of graphon W, often induces a particular latent structure amongst the vertices. Examples of such structure include partitions, latent features or hierarchies and we will use this description to group the model and talk about partition-based models, latent feature-based models and so on. We stress this classification is only approximate. Notice the idea of classifying models according to their parametrization of (U_i) and choice of graphon W was inspired by Lloyd et al. [2013].

6.2.1 Exponential random graph-models

An important type of network models which are not well-described using the W_{ij} notation of eq. (6.1) are *exponential random graph models* (ERGMs) such as the p^* model [Duijn et al., 2004, Holland and Leinhardt, 1981]. In this approach, one considers a (vector-valued) function s of a network A and assumes a density

$$\textit{(ERGM:)} \qquad p(A|\theta) = \frac{1}{Z} e^{\theta^T s(A)}, \quad Z = \sum_A e^{\theta^T s(A)}. \tag{6.3}$$

Notice this model may be interpreted as a maximum entropy distribution and s typically counts the number of triangles or edges. Models of this form, owning to their flexibility in choosing s, has historically played a prominent role in the literature of random graphs [Erdös and Rényi, 1959, Frank and Strauss, 1986, Holland and Leinhardt, 1981]. Exponential random graph models have however also been criticized (cf. Handcock [2003]) for their inferential complexity and degeneracy which significantly hinders the ability to estimate and sample the parameters θ in eq. (6.3), see also the very recent study by Chatterjee et al. [2013] which, surprisingly, showed for many summary statistics s the realized models are nearly identical to the Erdös-Rényi model.

6.2.2 Block-type models

The infinite relational model Kemp et al. [2006], Xu et al. [2006] and the stochastic block model [White et al., 1976, Holland et al., 1983, Wasserman and Anderson, 1987] provides the canonical example of a block-type model. In these $U_i \in \{1, \ldots, K\}$ (where $K = \infty$ for the IRM) and the link probability is

$$\textit{(Block models:)} \qquad W_{ij} = \eta_{U_i U_j} \tag{6.4}$$

where the entries of η is i.i.d. Beta distributed. Several models has been proposed which in effect only impose additional restrictions to η. These include letting the off-diagonal elements take the same value ($\eta_{\ell m} = \eta_0$ for all $\ell \neq m$) , imposing community structure by ensuring $\eta_{\ell\ell} \geq \eta_{\ell m} + \gamma$ [Mørup and Schmidt, 2012] (*Bayesian community detection*), imposing $\eta_{\ell\ell} = \eta_1$ and $\eta_{\ell m} = \eta_0$ for all $m \neq \ell$ [Hofman and Wiggins, 2008], drawing elements of $\eta_{\ell m}$ from a distribution containing fixed atoms (ie. $P(\eta \in \cdot) = \sum_i w_i \delta_{\eta_i} + P_0(\cdot)$) to either obtain a clustering such as a CRP or a slap-and-spike-type prior of the interactions, ie. multiple entries in η share the same (slap) rate η_0.

In addition the model trivially extends to weighted graphs. Suppose $A_{ij} \in \mathcal{X}$ and consider the case where $\mathcal{X}$ is a more general space than $\{0, 1\}$ and let $\mathcal{F}$

be a distribution on $\mathcal{X}$ with parameters θ in a space Θ equipped with a prior distribution H:

$$\theta_{U_iU_j} \sim H(\cdot) \tag{6.5a}$$
$$A_{ij}|(U_h),(\theta_{\ell k}) \sim F(\cdot|\theta_{U_iU_j}). \tag{6.5b}$$

Denoting the corresponding densities by p, if the integral over θ in eq. (6.5) is analytically tractable we say the model is conjugate and the resulting model is no harder to sample than the standard IRM. The idea of replacing the observational distribution with some other distribution was to our knowledge first proposed by Mørup and Schmidt [2012] in the case of a Poisson-Gamma observational model. We also attempted this approach to relational modelling for the case where $\mathcal{X} = \mathbb{R}$, F corresponded to a normal distribution, and the parameters θ was the mean and variance and H a Normal-Gamma distribution [Murphy, 2007]. The resulting model, dubbed the *normal infinite relational model* (NIRM), can be found in the publications given *Modelling Dense Relational Data* [Herlau et al., 2012b] which also contains implementation details. The same technique was subsequently applied to a number of problems in knowledge structuring by my co-author Fumiko Glückstadt [Glückstad, Herlau, Schmidt, and Mørup, 2014, Glückstad, Herlau, Schmidt, and Morup, 2013a, Glückstad, Herlau, Schmidt, Mørup, Rzepka, and Araki, 2013c, Glückstad, Herlau, Schmidt, and Mørup, 2013b].

A problem with e.g. the NIRM is relational data often contain a large number of entries which takes the same value, most often 0. In this case a Poisson, or more acutely, normal distribution as the choice of observational distribution F may be unsuitable. A simple but very incremental idea is to replace the observational distribution by a two-step procedure where one first generate a matrix B_{ij} of the non-zero elements and then draw the weights from the appropriate observational distribution. In the notation of eq. (6.5) this may be written as

$$\eta_{U_iU_j} \sim \text{Beta}(\eta_0^+, \eta_0^-) \tag{6.6a}$$
$$\theta_{U_iU_j} \sim H(\cdot) \tag{6.6b}$$
$$B_{ij}|(U_h),(\theta_{\ell k}) \sim F(\cdot|\theta_{U_iU_j}) \tag{6.6c}$$
$$I_{ij}|(U_h),(\eta_{\ell k}) \sim \text{Bernoulli}(\eta_{U_iU_j}) \tag{6.6d}$$
$$A_{ij} = B_{ij}I_{ij} \tag{6.6e}$$

If a sample from F is zero with probability zero, for instance if F correspond to a normal observational model or one plus a Poisson distributed random variable, one can integrate out η and θ and this model will likely be more suitable for data with many zero entries.

6.2.2.1 Block-type models with weights

The final simple extension to block-type models is the use of weights. Networks often contain vertices of vastly different degree but similar community structure. For instance in a social network two people (an introvert and an extrovert) may have the same interests and form friendships along the same community structure, but the extrovert may simply form *more* friendships. This is the idea behind the *degree-corrected stochastic block model* [Karrer and Newman, 2011, Newman, 2012] which admits a representation

$$\textit{(DCSBM:)} \qquad W_{ij} = U_i' U_j' \eta_{U_i U_j} \tag{6.7}$$

where $U_i \in \{1, 2, \dots\}$ indicate community membership and (U_i') is a list of scalars indicating the *weight* of vertex i. The observational distribution is Poisson with rate W_{ij}. In terms of inference, Karrer and Newman [2011] considered an optimization scheme; in the included work *Infinite-degree-corrected stochastic block model* [Herlau et al., 2014b] we considered a Bayesian formulation of eq. (6.7) which admitted both the weights (U_i) and interactions η to be integrated out, specifically for $i \neq j$ and assuming a partition $\pi = \{B_1, \dots, B_k\}$ where $k_\ell = |B_\ell|$:

$$\eta_{\ell m} \sim \text{Gamma}(\lambda_a, \lambda_b) \tag{6.8a}$$
$$(\theta_\ell) \sim \text{Dirichlet}((\gamma)_{i=1}^{k_\ell}) \tag{6.8b}$$
$$W_{ij} = k_{U_i} k_{U_j} \theta_{iU_i} \theta_{jU_j} \eta_{U_i U_j} \tag{6.8c}$$
$$A_{ij} \sim \text{Poisson}(W_{ij}) \tag{6.8d}$$

where U_i denotes assignments to blocks in the partition π distributed as a CRP. The resulting model leads to a sampling scheme no more costly than the IRM, however this comes with the cost of not being exchangeable.

6.2.3 Distance and norm-based models

A slightly more general approach than partition-based models is consider each U_i as points in a vector space (typically $\mathbb{R}^h$ for a small integer h) and a representation

$$\textit{(Latent distance based models:)} \qquad W_{ij} = -d(U_i - U_j) \tag{6.9}$$

for a distance measure d typically the euclidian distance. Since W_{ij} is negative it should then subsequently be put through a linear transformation followed by a logistic map. This type of model is called a *latent space* based model [Hoff et al., 2002] and U_i can be interpreted as the position of a vertex in a *social*

space [Tang and Liu, 2009]. Thus, for $h = 2$ model may be considered a visualization approach and as such share important conceptual features with Kohonen projections [Kohonen, 1988].

An approach very related to this model is the *eigenmodel* [Hoff, 2007] where the observational model is

$$\textit{(Eigenmodel:)} \qquad W_{ij} = -U_i^T D U_j \tag{6.10}$$

for a diagonal matrix D. Notice if the distance d in eq. (6.9) is in an inner-product space with inner product $\langle \cdot, \cdot \rangle$ we have the identity $d(U_i, U_j) = \langle U_i, U_i \rangle + \langle U_j, U_j \rangle - 2\langle U_i, U_j \rangle$ and so for positive semidefinite matrices D the latent space model eq. (6.9) and eigenmodel eq. (6.10) are very similar. We will return to the case of models with the same basic algebraic structure in the next section.

6.2.4 Latent feature-based models

Latent-feature based models are models in which the underlying structure is interpreted as each vertex i having access to multiple features. The canonical example is a social network where each vertex consists of a person and the edges to friendships and the features may be workplaces, schools, a particular football club or broader features like sex or age group. Each vertex may then select zero or more features and these features, and we write this assignment as $U_i \in \{0, 1\}^K$ where K may be ∞, assuming each U_i only contain a finite number of non-zero elements. The *latent feature relational model* of Miller, Griffiths, and Jordan [2009], Miller [2011] is then written as

$$\textit{(LFRM:)} \qquad W_{ij} = U_i^T D U_j \tag{6.11}$$

where a normalization procedure is again implicit and D is assumed to be a general matrix typically with i.i.d. normally distributed entries. As a prior for the set of feature-assignments (U_i) an obvious choice, assumed in the following, is the *Indian buffet process* (IBP) [Griffiths and Ghahramani, 2005] however we will not discuss the details here.

A model falling somewhere between the LFRM and the partition-based models described previously is the *Mixed-Membership stochastic block model* [Airoldi et al., 2008]. In this model, the Bernoulli edge rate can roughly be written as

$$\textit{(MMSBM:)} \qquad W_{ij} = U_i^T \eta U_j \tag{6.12}$$

where η is a symmetric matrix of i.i.d. beta-distributed entries and, as opposed to the LFRM, each U_i belong to the K dimensional unit simplex $\{x : x_k \geq$

0 and $\sum_k x_k = 1\}$ and has the interpretation of selecting a *mixed* membership between K clusters.

The formulation of the LFRM given in eq. (6.11) is very general and negative (diagonal) entries of D implies having the same feature reduce the chance of an edge. For this reason, and to simplify the inference procedure, a more structured interaction between the features give the *Infinite Multiple Membership Relational Model* [Morup et al., 2011]

$$\textit{(MMRM:)} \qquad \log(1 - W_{ij}) = \log(1 - \eta_0) + U_i^T \log(1 - \eta) U_j \tag{6.13}$$

where each $\eta_{\ell k}$ is assumed Beta distributed, $\log(1 - \eta)$ is considered a matrix where entry ℓk is $\log(1 - \eta_{\ell k})$ and W_{ij} is the link probability, i.e. the parameter of a Bernoulli random variable as in eq. (6.1). The interpretation of the model is that sharing features always increases edge probability.

An extension of the latent feature relational model is the *infinite latent attribute model* [Palla et al., 2012] in which, in addition to assigning zero or more features to each vertex i given by $\tilde{U}_i \in \{0, 1\}^\infty$, for each feature m a partitioning of the vertices is generated $V^{(m)} \in \{1, 2, \ldots, \infty\}^\infty$. The generative model can now be written as

$$\textit{(ILA model:)} \qquad W_{ij} = \sum_m \tilde{U}_{im} \tilde{U}_{jm} D^{(m)}_{V_i^{(m)} V_j^{(m)}} \tag{6.14}$$

where each $D^{(m)}$ is an infinite matrix with i.i.d. normally distributed entries. The model may also be written more compactly, introducing $U_{im} \equiv \tilde{U}_{im} V_i^{(m)}$, as

$$W_{ij} = \sum_m 1_{U_{im}} 1_{U_{jm}} D^{(m)}_{U_{im} U_{jm}} \tag{6.15}$$

where 1_k is the indicator function equal to 1 iff. $k > 0$. Comparing to the LFRM eq. (6.11), the main difference is each feature, U_{im}, is subdivided into one or more non-overlapping subfeatures. For instance a feature *play sports* may contain the subfeatures *golf*, *soccer*, *tennis* and so on which cannot overlap.

Such a hierarchical extension is not trivially the most appropriate ontology. Consider for instance the case of the sex of a person; one can either consider it as a feature (*the sex*) with two subfeatures (*male* or *female*) with the problem a person may have no sex at all or alternatively, one can consider two top-level features *sex-male* and *sex-female* with a single subfeature with the problem a person can then be both male or female). However the ILA is to our knowledge the Bayesian model which offers the best link prediction. A disadvantage of the ILA is the hierarchical composition of two discrete data structures, a feature assignment and a partition, makes scalable sampling a difficult problem.

6.2.5 Continuous feature-based models

An assumption the above models has in common is that the feature assignments are discrete. This assumption can be relaxed by letting U_i be a general vector in $\mathbb{R}^h$ for some h; we encountered this structure in the previous section for the latent-distance and eigenmodel, however in terms of modelling goal and interpretation of each coordinate of U_i it is commonly thought of as a latent feature. This interpretation is particularly tempting for the case where $U_i \in \mathbb{R}^h_+$ which will be discussed later. The advantage of this formulation is the matrix of interaction, for instance D for the LFRM eq. (6.11), may be absorbed into the latent feature matrix U_i leading to a simplified formulation.

One particular example is the *probabialistic matrix factorization* method of Mnih and Salakhutdinov [2007]. Letting $\sigma(x) = (1 + e^{-x})^{-1}$ denote the sigmoid function and assuming $U_i, V_j \in \mathbb{R}^h$ two general vectors this model rely on a representation

$$\textit{(PMF-based model:)} \qquad W_{ij} = \sigma\left(U_i^T V_j\right) \tag{6.16}$$

where W_{ij} is thought to represent the *mean* of a normal distribution for *weighted* bipartite graph data. The simplicity and choice of normal prior lead to an efficient inference method and the paper contain several interesting details on handling missing data making it highly recommendable. The same basic algebraic form of PMF was given a throughout treatment by Menon and Elkan [2011, 2010]. Though the focus is on scalable inference in a likelihood-maximization framework and integration of background information the network-specific parts of the model may be well be formulated as

$$\textit{(Menon-Elkan model:)} \qquad W_{ij} = f\left(U_i^T V_j + a^T U_i + b^T V_j + c_0\right) \tag{6.17}$$

where different choices of link-function f is considered for instance the sigmoid function for binary data. In the above a, b are h-dimensional vectors and c_0 is a bias term.

Both the PMF and the Menon-Elkan approach in eq. (6.16) and eq. (6.17) share some common traits with matrix decompositions such as a SVD or PCA and, in particular when the coordinates are given a latent-feature interpretation, with *non-negative matrix* factorization [Lawton and Sylvestre, 1971, Paatero and Tapper, 1994]. Since we are interested in Bayesian models we will certainly not attempt to review this vast literature but only review selected Bayesian incarnations. The simplest is Bayesian non-negative matrix factorization [Cemgil, 2009], in this model for bipartite relational data the observations are assumed to be Poisson distributed with a rate W_{ij} given by

$$\textit{(BNMF:)} \qquad W_{ij} = U_i^T V_j \tag{6.18}$$

and $U_i, V_j \in \mathbb{R}_+^h$, the most obvious choice being an i.i.d gamma distribution. Notice the choice of a Poisson-gamma parametrization lead to analytical simplifications. The correlated topic modelBlei and Lafferty [2006] and latent Dirichlet allocation Blei et al. [2003] share very similar decompositions, see for instance Paisley et al. [2014] for a modern review and discussion of large-scale implementation using stochastic variational Bayesian methods. If we break the Aldous-Hover representation by admitting correlation between the U_i vectors many other matrix factorization methods may be given a probabilistic interpretation [Shashanka et al., 2008, Singh and Gordon, 2008].

6.2.6 Random-Function based models

Since the entire discussion has been focused around the Aldous-Hoover theorem in which the central object, the graphon, *is* a random function the title is perhaps a bit misleading. What is being referred to is that the graphon is generated "directly" and not derived from other structural assumptions such as feature allocations or partitions. An important example is the Mondrian process of Roy and Teh [2008]. The Mondrian process is a prior of piecewise constant functions over a rectangular domain. One way to describe the Mondrian process is by the following recursive procedure to sample a function W on some domain: For a rectangular domain $B_1 \times B_2$, B_1, B_2 denoting intervals of $\mathbb{R}$, either (i) let the function W take a constant value x sampled from some distribution (for instance a Beta distribution) on $B_1 \times B_2$ or (ii) select either B_1 or B_2, say B_1, and subdivide B_1 into two new intervals $B_1 = b_1 \cup b_2$ and recursively call the Mondrian process on two rectangular domains $b_1 \times B_2$ and $b_2 \times B_2$. Naturally many variations of this procedure exists, see Roy and Teh [2008] for additional information. Generatively we might write a Mondrian-process based model for uniformly distributed random scalars $(U_i)_i$ as

$$\textit{(Mondrian-based model:)} \qquad W \sim \text{MondrianProcess}(\cdot) \tag{6.19a}$$

$$W_{ij} = W(U_i, U_j). \tag{6.19b}$$

The second idea, first proposed by Lloyd, Orbanz, Ghahramani, and Roy [2013] is to use a composite function of a sigmoid function and a Gaussian process as prior for W:

$$\textit{(GPRM:)} \qquad W \sim \text{GP}(0, \kappa) \tag{6.20a}$$

$$W_{ij} = W(U_i, U_j). \tag{6.20b}$$

The better predictive performance for GPRM is obtained when the random elements U_i belong to a vector space such as $U_i \in \mathbb{R}^4$. Notice our description omits many important details such as the appropriate symmetrization of kernel function and the use of an alternative parametrization to obtain reasonable scalability. [Lloyd et al., 2013]

6.2.7 Random hierarchy-based models

Just as vertices can be divided into partition to give rise to partition-based models such as the IRM, or vertices can be assigned features vectors to give feature-based models such as the LFRM, vertices can be organized into a hierarchy to give rise to hierarchical models. Generally speaking, a hierarchy may be considered a feature assignment - the features associated with each vertex being the path towards the parent node in the hierarchy. As a consequence, a random hierarchy-based model can be considered more flexible than a random partition-based model but less flexible than a random feature-based model.

As we have extensively reviewed hierarchical models in the papers Herlau et al. [2012a] and Herlau et al. [2013] we will only give a brief review here assuming the notation introduced in section 4.4. Suppose for each vertex i, the symbol U_i denote a *node* in a random hierarchy t. Consider two nodes n_1 and n_2 in t such that n_1 is not a descendant of n_2 and n_2 is not a descendant of n_1. In this case there is a unique node n of the tree such that

- n is a parent of m_1 and m_2
- n_1 is a descendant of m_1
- n_2 is a descendant of m_2

This is simply saying that starting from two nodes n_1 and n_2 and tracing up in the hierarchy towards the root we will eventually encounter a common node n. We will denote the previous operation by the function

$$f(n_1, n_2) = \begin{cases} \{m_1, m_2\} & \text{if } n_1 \neq n_2 \\ \{n_1\} & \text{otherwise} \end{cases} . \tag{6.21}$$

Notice the convention if $n_1 = n_2$ the function f simply returns the current node and $f(n_1, n_2) = f(n_2, n_1)$.

The simplest hierarchical model, the *binary hierarchical relational model* of Clauset, Moore, and Newman [2008], can the then be defined as the generative procedure where each $U_i = \{i\}$ correspond to the leafs of a *binary* random hierarchy t and

$$\textit{(bHRM:)} \qquad W_{ij} = \eta_{f(U_i, U_j)} \quad \textit{(t binary)} \tag{6.22}$$

where each η-value is Beta distributed and W_{ij} parameterize a Bernoulli distribution. Clauset, Moore, and Newman [2008] considers a uniform prior over the binary hierarchy t. In the given notation this model can easily be extended by

simply letting $U_i = i$ as before correspond to the leaf set of t but this time let t be a *general* Gibbs fragmentation tree of the type considered in section 4.4.1. This model, dubbed the *hierarchical relational model* (HRM), was considered by us in Schmidt, Herlau, and Mørup [2013] and a similar idea for continuous data in Schmidt, Herlau, and Mørup [2014]. A potential drawback of a hierarchical model is the hierarchy is possibly not meaningful at a certain level. We considered a simple extensions of the HRM in which $U_i \in \{1, 2, \dots\}$ (the distribution of (U_i) being a CRP) and consider t as a tree with leaf-set again given by the unique values of (U_i). To put this procedure slightly differently, the model first partitions the vertices according to a CRP, then build a tree where each leaf corresponds to exactly one block in the partition and proceed as in eq. (6.22) giving:

$$\textit{(HRM, hierarchical IRM:)} \quad W_{ij} = \eta_{f(U_i,U_j)} \quad \textit{(t general hierarchy).} \tag{6.23}$$

Details can be found in *Detecting Hierarchical Structure in Networks* [Herlau et al., 2012a]. A variation of this method based on fixed-depth hierarchies is also discussed by Ho et al. [2012].

A common theme for the HRM and hierarchical IRM is the structure of the hierarchy is used to *directly* generate the parameters $\eta_\cdot$ of the observational distribution. The advantage of this approach is the η's may be integrated out and inference reduce to sampling a single discrete structure which can be accomplished quite efficiently. More sophisticated approaches forego this advantage but gain a more flexible construction. An important example of this type of model is that of Roy, Kemp, Mansinghka, and Tenenbaum [2007] who considers (the model appears not to be named in the paper) the *hierarchical Mondrian relational model* (HMRM) which can be thought of as the Mondrian process based relational model with the important distinction the (recursive) subsplits are not independent. To be more exact, consider the unit interval $[0, 1]$ and a recursive partition structure. That is, an infinite binary tree t such that the root of t correspond to the unit interval, each node n to a subinterval B_n of $[0, 1]$ and such that for any node n of t with children n_1, n_2 we have that B_{n_1}, B_{n_2} is a partition of B_n.

The hierarchical Mondrian process may now be described as a random function model for a function W on the unit interval by the following procedure initialized with (n, n) where n is the root node of the hierarchy: Given two nodes (n, m) with some probability, either (i) let W take a random value from a fixed distribution on $B_n \times B_m$ and terminate or alternatively, (ii) select one of the nodes, say n with children n_1, n_2, and recursively call the method two times as (n_1, m) and (n_2, m). To be explicit:

$$\textit{(HMRM:)} \quad W|t \sim \text{HierarchicalMondrian}(\cdot|t) \quad \textit{(t binary).} \tag{6.24}$$

The above description leave out many important details including if the method converge at all. Furthermore the authors show the above procedure can be given a discrete parametrization which admits the entire recursive procedure to be integrated out such that only t need be sampled [Roy et al., 2007]. Evidently, the process may be extended in several ways such as to include non-binary hierarchies.

A feature of the HRM and in particular the HMRM is despite the ability to integrate out many of the variables the inference method is still quite computationally intensive. This is due to the number of operations required to compute the change in likelihood resulting from a change in a hierarchy tend to scale in the number of nodes in the hierarchy compared to the number of blocks in a partition; furthermore each of these operations often have a higher (constant) cost.

Recently, Blundell and Teh [2013] proposed *Bayesian hierarchical community discovery*. This method attempts to alleviate the computational cost while maintaining flexibility using a greedy agglomerative clustering technique to construct the underlying hierarchical structure. The structural assumptions in the model may (very roughly) be described as an interpolation between the HMRM and the HRM, however the method is very scalable and provides surprisingly good link prediction.

6.3 Temporal Models

A *temporal networks* is a network which changes over time. We will distinguish between the following cases: The first is networks where the edges are *temporally persistent*, that is, the edges persist over intervals of time. Examples could be friendships in a social network or the physical connectivity of computers on the internet. The second is networks where the edges are *temporally intermittent* meaning the edges denotes instantaneous interactions. Examples could be social interactions (relative to a time scale of days) or electronic communication on the internet.

For both types of networks one may observe actual snapshots of the network, that is, for a persistent network the states of the edges at various time points or for a temporally intermittent network a point set of events of the form $(\text{Vertex}, \text{Vertex}, \text{Time})$ or alternatively one may observe the aggregated network at different time points. The friendships in a digital social network are often good examples of an aggregated network since edges are very rarely deleted even if they no longer correspond to an ongoing social interaction.

In addition to these effects it is often (but not always) appropriate to consider the vertices as having different life spans. A slightly morbid example is a friendship network where people are born and die.

The distinctions illustrate that *temporal network data* do not refer to any one single thing and temporal models suitable for *one type* of data can be quite unsuitable for other types. This is not to say we consider all non-temporal (stationary) network data as something best treated by any single network model, however the problem is far more obvious for temporal networks and the space of potentially models is both much larger and less explored than for stationary networks.

We will therefore not attempt to give a very throughout review but only mention a few examples of important work and give a general overview of the field. Firstly, for practical reasons, we will explicitly assume temporal network data only refer to snapshots of a network $A_1, \ldots, A_T$ at discrete times and assume vertices may have limited time spans or not depending on the situation.

6.3.1 Examples of temporal models

With these preliminary remarks, temporal models for network data are typically obtained by taking an existing stationary model and adding a temporal effect to some part of the parametrization. In the Aldous-Hoover inspired notation this may either be to the latent vertex properties U_i, to the interaction-function W or to both. There are two notable exceptions to this program, the first being temporal versions of the exponential random graph model by e.g. Robins and Pattison [2001], see also Guo et al. [2007] for a ERGM-based approach to inferring rewiring dynamics. The second example is the proposal of Heaukulani and Ghahramani [2013]. The proposed model share similarity with the LFRM of Miller et al. [2009] in terms of parametrization, however with the novel proposal it is the *actual realization* of the network at time t which affect the network at time $t + 1$ and not (only) the underlying latent parametrization at time t. Whether this choice allows significantly better link prediction requires more exploration of the model space, however the assumption is no doubt more accurate in situation where the edges really denote channels of interaction.

Returning to models conceptually closer to temporal extensions of existing stationary models some of the more notable examples are, for the IRM, the dynamic extensions using either a discrete hidden Markov model based approach [Ishiguro et al., 2010] or Kalman-filters [Xu and Hero III, 2013]. Latent-space based models are perhaps easiest amendable to a dynamical formalism by, for instance, applying a Gaussian process to give temporal correlation on the latent embed-

ding. Approaches roughly falling within this framework include the original work on latent embedding Hoff et al. [2002] and ideas based on survival analysis where the problem of inferring the temporal network is considered a regression problem based on time-dependent network statistics and temporal regression coefficient [Vu et al., 2011, Perry and Wolfe, 2013]. This approach has been revisited several times, see for instance the dynamic latent-space model of [Sarkar and Moore, 2005] and, by parameterizing the mixed-membership indicator variable U_i for the mixed-membership stochastic block model in eq. (6.12) as a transformed normal variable, the dynamic mixed membership stochastic block model [Fu et al., 2009]. See also the Gaussian-process based approach of Durante and Dunson [2014] for a principal approach to dynamic embedding within a Bayesian framework based on Gaussian processes.

For discrete feature-assignments such as the LFRM a basic problem is dynamic evolution of feature assignments are quite hard to evolve in a principal manner. Two approaches bear mentioning, one is dynamic infinite relational latent feature model (DRIFT) [Foulds et al., 2011] which proposes a dynamical Markov process on the feature assignments (but keep the interactions fixed). A possibly more principled approach is to study a probabilistic object corresponding to a time-evolving (or time-dependent) feature assignment. This approach was taken by Leskovec [2013] which relied on the distance-dependent IBP [Gershman et al., 2011].

6.3.1.1 Temporal hierarchies

An obvious idea at this point is to consider a temporal extension of hierarchical relational models. How to proceed with this program is somewhat non-obvious in that there to our knowledge do not exist well-explored temporal (or dependent) priors for Gibbs fragmentations. We attempted to overcome this problem by relying on a construction in which the vertices could be grouped together across time (implying they did not change role) and the unique states of the vertices obtained in this manner was gathered in one single hierarchy distributed as a Gibbs fragmentation tree. By virtue of marginal consistency of a Gibbs fragmentation tree, one can show this construction induces temporally dependent fragmentation.

While this procedure has some merits, in particular that the marginal distribution at each time slice is distributed as a Gibbs fragmentation tree, analytical simplicity and the ability to share parameters across time (which we nevertheless did not explore), we do not consider the proposal to be a "true" temporal process on hierarchies (in the same sense the distance-dependent IBP of [Gershman et al., 2011] is a true temporal process on feature assignments) for the reason

the chain of networks are either temporally uncorrelated or have infinite memory. Finding a computationally tractable structure without these deficiencies is to our knowledge still an open problem. In our article *Detecting Hierarchical Structure in Networks* [Herlau et al., 2013] we provide a full account of the THRM from a generative perspective, discuss its properties and evaluates it on three temporal network data sets.

Chapter 7

Discussion and Conclusion

During the past chapters I have tried to give a brief account of the ideas the written work relies upon and relates to. In the discussion, I will give my view of which challenges I believe are the more interesting and, where applicable, use this to give a critical treatment of the written work. I have deliberately chosen to structure the discussion in reverse order of how the topics are laid out in the main text.

Bayesian methods for networks: Concerning Bayesian methods for networks I believe there are two robust observations to be made. The first is complexity matters. There appears to be distinct gains in performance when going from the IRM (a partition) to the LFRM (a feature allocation) to the ILA model (both).

The second is exact inference (i.e. asymptotically exact such as MCMC samplers) of Bayesian models for networks is very computationally intensive and appears to put upper limits on the system sizes which can plausibly be sampled. For the IRM it was my experience this limit may be as low as 300-600 vertices using the considered methods [Herlau et al., 2014a], and this upper limit does not seem to be increasing very fast. This suggest one should place more emphasis on approximate inference methods such as variational methods for the more advanced network models which offer the better predictive per-

formance. Variational methods, in particular stochastic variational Bayes, has been applied extensively to the LDA and non-negative matrix factorization-like methods [Paisley et al., 2014], however for more advanced models in the context of *network* data, stochastic variational Bayes has to my knowledge not yet been widely applied for models more complicated than block-type models. It should of course be kept in mind there may be good reasons for this, perhaps the resulting method is not tractable or, alternatively, it simply fails to perform well.

In my opinion, the space of Bayesian network models of the particular *form* suggested by the Aldous-Hoover representation appears fairly well-explored at this point and it is in this respect important to focus on some limitations inherit to this particular representation. I have previously discussed one of these, namely sparsity. Models derived from the alternative Poisson process representation [Kallenberg, 2005, chapter 9] I discussed in section 4.3.1.1 avoids this problem and in my opinion they could represent one of the more significant contributions made to network modelling in the past years. It is important to emphasize the existing work on applying this representation to networks by Caron and Fox [2014] only models vertex degree. Models based on this representation but also exploring community structure, feature allocations or the temporal evolution of networks is to my knowledge an area fully open for exploration and it would seem a safe bet this work will be undertaken in the coming years.

A second important limitation of an Aldous-Hoover type representation is the control on structural motifs such as triangles. Some care is required to even formulate this problem, for instance a network which contain no edges outside a dense subgraph is organized to contain a maximum number of triangles relative to the number of edges, however this is quite evidently not what we typically intend when we consider networks with *"many triangles"*. Interestingly, the point-process representation of Caron and Fox [2014] may offer some helpful guide to constructing sparse networks with an over-representation of triangles. Let Π be a draw from a Poisson point-process on $\mathbb{R}^3$ and assume (for simplicity) the base measure is the uniform measure. If we interpret each point (x_i, y_i, z_i) in Π as a *triangle*, that is, as saying there are edges $x_i y_i$ and $y_i z_i$ and $x_i z_i$ in the graph, the underlying infinite graph will quite obviously contain far more triangles than could be expected from its other properties.

When considering this problem a pointed question is *why* a graph contains many triangles. In many cases the answer seems tightly tied to a temporal dynamic, say, two people with a common friend are likely to be introduced to each other at a later point. In this case one should perhaps not simply seek a model with an over-representation of triangles for a single network, but one of plausible social dynamics *which gives rise* to triangles such as the latent feature propagation

model [Heaukulani and Ghahramani, 2013].

One striking feature of network modelling is the method used to obtain the network can play a more crucial role than the underlying structure. Consider two methods of registering a large social network of 10^6 people: one being a breadth-first search starting from a single vertex and proceeding to obtain 1000 vertices, the other being selecting 1000 people at random and reporting their subgraph. These datasets will differ in every respect, yet in terms of motivation of a particular latent structure for network data and subsequent analysis, the network is commonly assumed as simply being what it is. While this problem has been the focus of some work [Mislove et al., 2007, Kurant et al., 2010, Ferrara et al., 2012] the problem is often simply ignored; My own work is certainly not an exception to this practice.

With respect to hierarchies in relational modelling, one lesson seems to be the use of a single hierarchy as organizational principle such as in the HRM or THRM will lead to inferior link prediction compared to e.g. the LFRM or ILA model. Another issue is, while most network data will exhibit hierarchical features, any single hierarchy of the vertices will be insufficient. Solutions to this problem may include the use of several hierarchies on the vertices; however this proposal would inevitably come at the cost of interpretability. An additional challenge is that building efficient samplers for hierarchies became a time-consuming pastime during this thesis, both in terms of implementation time and subsequently when the samplers were evaluated. Samplers for hierarchies may *theoretically* have good scaling properties (see for instance the discussion in Schmidt, Herlau, and Mørup [2013]), however in practice they appear to be quite difficult to scale to larger networks. We are very tempted to think this will remain a problem for any model which attempts to infer a single hierarchical structure over *all* elements. Tackling this problem may include building many smaller hierarchies or, as in the hierarchical community discovery model [Blundell and Teh, 2013], consider a Bayesian approach build on top of hierarchies that are not directly sampled.

Sampling For a long time during completion of this thesis I considered sampling a secondary concern. I recall both thinking and saying the advantage of split-merge sampling over Gibbs sampling was small at best for the IRM. Evidently I now consider myself to have been wrong on both counts. No doubt some of my confusion on this point was my own fault, however the wider Bayesian literature, and non-parametric modelling of networks in particular, often treats sampling in a somewhat strange manner. Most work contain a detailed description of the sampler, often because the sampling scheme is elaborate and the proposal contains several ingenious ideas, however it is very often the case

no results are presented on the convergence of the sampling scheme and no comments are made if unreported experiments suggested the sampler reached equilibrium.

At least for blockmodels there are good reasons to think many samplers do *not* reach equilibrium as described in Herlau et al. [2014a]. While one could argue a lack of mixing might not matter much from a pragmatic perspective, one cannot help but think the mixing properties is an aspect of non-parametric Bayesian modelling which should play a greater role.

As to my own suggestion in this respect it is in the main a matter of the common-sense intuition samplers should allow as large a move class as possible and some engineering. The aspect of the proposal which I consider the most important is that of using other states to construct proposals. That is, if we have a system with property x, and we wish to change that property to y (during a proposal move), having two other systems, one with property x and another with y, will offer important clues as to what other aspects of the first system needs to be altered to allow property x to change to y. For discrete systems I consider this to be an idea well worth exploring for feature allocations or hierarchies.

Non-parametrical methods Most Bayesian models contains parameters corresponding to a number of clusters, a degree of a polynomial, a number of features or a depth of a tree. If we adopt the qualitative and very broad sense of the word non-parametric as referring to models where these parameters are inferred it seems quite self-evident one would have to throw out a great deal of the justification for the use of probabilities to insist one should not put priors on these parameters and treat them as other unknown parameters of the model. I will therefore leave aside this definition and consider three other ways to envision non-parametrics and its relationship to Bayesian modelling.

Firstly, one can consider non-parametrical methods as certain general classes of priors for data structure. This is naturally tied into the above description. If one has a model based on partitions the partition can be thought of as distributed as a CRP, if one has a model with latent features one can consider an IBP and if one has a model which makes use of a function one can consider a prior based on a Gaussian process, a Mondrian process or a jump process based on a beta process. For a long time I implicitly considered Bayesian non-parametrics in this fashion and, provided one keeps the statistical properties of the various processes in mind, it is a fruitful way to "infinitize" various models. A danger of this view is to miss the greater picture, namely that non-parametrical methods in the mathematical statistics literature arises from invariances.

Thus, the second approach to non-parametrics is to think of any model as something which has to obey a particular invariance related to the data type. For instance a model of networks *should* be doubly exchangeable and exchangeability is then in and by itself a major selling point of the model. While invariance often automatically ensures graceful behaviour when data is missing, see the example of the robot in chapter 4, an assumption of exchangeability is fundamentally an assumption on the data-generating process and as such it can be right or wrong and may serve as a strait-jacket when applied uncritically.

The third way to be influenced by non-parametrical methods is by the observation the mathematical literature of probability theory provides very general ways of identifying the underlying structure in a particular problem through representor theorems, for instance the Aldous-Hoover-type parametrization of W_{ij} as $W(U_i, U_j)$ and the interpretation of (U_i) and W as random objects. This identification goes beyond if the resulting model is actually exchangeable –this was not the case for many of the examples in section 6.2– and in our opinion allows results in the mathematical statistics literature to influence creative thinking in machine learning in a way that goes beyond the "infinitizing" described above. While we can only guess how people are influenced in practice we would like to draw attention to the work on sparse graphs by Caron [2012] and the Aldous-Hoover representation by Lloyd et al. [2013].

Updating beliefs, use of probabilities and outlook In my opinion, probabilistic or Bayesian methods are important in machine learning for three reasons. Firstly, for many problems Bayesian methods offers very good modelling performance and even when they do not they tends to be robust. Secondly, Bayesian methods rest squarely on a mature mathematical field, probability theory and thirdly, Bayesian methods can be motivated by analyzing what appears to be a mental phenomenon – beliefs.

From the later perspective it is tempting to consider machine learning as an attempt to derive a theory of thinking. What, then, can be expected to bring about progress? Machine learning as a field is still so young that it is very difficult to single out what will stand back as lasting accomplishments 50 years from now. To put this in familiar terms, the dataset is too small, too noisy and too poorly labeled. However the view that machine learning is a theory of thinking makes the analogy to physics more apt, and physics has the advantage of 400 years of progress and a well-documented history. If we admit this as training data, perhaps a few more things can be said.

Reflections on the progress of physics often takes the view progress consists of two parts. One is an accumulation of empirical facts, the other is smaller

or larger re-organizations of the theoretical picture in which discrete (and significant) changes to the vocabulary and assumptions both re-organize and reinterpret the empirical facts to fit the new overarching understanding. From the perspective of different scientific models what constitutes basic terminology differ to the extent two representatives of different scientific pictures cannot view each other's ideas from an objective viewpoint.

This view, especially when not stated as crudely as above, certainly represents some aspect of the relationship between scientific theories. But undue emphasis risks obscuring an important aspect of scientific progress namely the *conservative* aspect of scientific progress. I will illustrate this with a few examples. In the late part of the 19th century, there was an obvious contradiction between the Galilean transformations and Maxwells equations. The Galilean transformations suggests all inertial frames are equivalent, however in the Maxwell equations a stationary charge and a non-stationary charge differ in that the stationary charge has no magnetic field while a moving charge do. To avoid this contradiction one could consider two approaches. One could do away with Gallilean invariance, that is, accept some inertial systems are not equivalent with respect to electromagnetic phenomena, or alternatively, consider Maxwells equations to be an approximation of a more elaborate theory of wave-phenomena in an ether.

Einstein's contribution was exactly to recognize one did *not* have to make new theoretical innovations along either path. Einstein *accepted* Maxwell's equations sans ether and that Galileos insight, that inertial systems are equivalent, simply could not be wrong. What he realized was these two only led to a contradiction in the context of other ideas, most importantly that equivalence of inertial systems and the Galilean transformation are not the same thing and from this followed the theory of relativity.

The same pattern can be found in other scientific innovations: If one accepts general relativity as the theory of gravity and the observation the universe is isotropic one arrives at a dynamic universe. Newton famously recognized the central concept of dynamics was acceleration and the suns role in Keplers dynamics had to be explained through acceleration. If one accept electromagnetism is propagated by fields but the world must be quantized one arrive at quantum-field theories as Dirac famously discovered.

It is easy, indeed, tempting to look at these innovations from the viewpoint what they changed, but again and again it seems what is more important is which concepts the inventors insisted *could not* be subject to change and what other concepts *were independent of these*.

If we return to machine learning, clearly no aspect of the previous discussion contains a formula for success. Understanding what elements of current ideas

should be taken serious and which should be taken as spurious requires a Newton, Einstein or Dirac. It is however in my opinion an interesting observation that Bayesian methods can be split into a consistency requirement and an invariance principle that allow the assignment of beliefs. It should be noted My own work on making use of this distinction [Herlau et al., 2015] has only led to modest results. Are there other aspects of Bayesian learning that may be separated? If we ask what Bayesian learning is according to chapter 2, it is a semantic of belief on propositions which has the property of truth and falsehood. If we ask a linguist he will wonder why we do not care about vagueness. If we ask a philosopher of the Wittgensteinian school he may say our propositions has the property of truth and falsehood only because we choose to formulated them in a particular way and if we ask a neuroscientist he might wonder what image of the mind this suggest as dynamics play no part of the theory.

The most important innovations in science have not been based on radical postulates but either on specific observations or a conservative and almost pedantic investigation of existing ideas. Nearly everyone I have spoken to in my field agrees we need radically new insights to explain intelligence. This may be the case, but from the vantage point of history it is perhaps more worthwhile to look at what existing insights are good candidates to *remain* relevant. It is in this sense I think the Bayesian view on beliefs has an important future role to play.

Bibliography

J. Aczél. *Lectures on Functional Equations and Their Applications*. Mathematics in science and engineering. Academic Press, 1966. URL `https://books.google.com/books?id=0vZQAAAAMAAJ`.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(1981-2014):3, 2008.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

David Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198, 1985.

David Aldous. Probability distributions on cladograms. In *Random discrete structures*, pages 1–18. Springer, 1996.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

David J Aldous. Exchangeability and continuum limits of discrete random structures. In *Proceedings of the International Congress of Mathematicians*, volume 1, pages 141–153, 2010.

Romas Aleliunas. A Summary of a New Normative Theory of Probabilistic Logic. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '88, pages 199–206, Amsterdam, The Netherlands,

The Netherlands, 1990. North-Holland Publishing Co. ISBN 0-444-88650-8. URL http://dl.acm.org/citation.cfm?id=647231.719560.

SI Amari. Neural learning in structured parameter spaces-natural Riemannian gradient. *Advances in neural information processing systems*, pages 127–133, 1997.

Karen S Ambrosen, Tue Herlau, Tim Dyrby, Mikkel N Schmidt, and Morten Morup. Comparing Structural Brain Connectivity by the Infinite Relational Model. In *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*, pages 50–53. IEEE, 2013.

Kasper Winther Andersen, Tue Herlau, Morten Mørup, Mikkel N. Schmidt, Kristoffer H. Madsen, Mark Lyksborg, and Hartwig Siebner. Joint modelling of structural and functional brain networks. In *NIPS workshop on Machine Learning and Interpretation in Neuroimaging*, 2012.

Christophe Andrieu and Christian P Robert. *Controlled MCMC for optimal sampling*. Citeseer, 2001.

Christophe Andrieu, Éric Moulines, et al. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462–1505, 2006.

Takamitsu Araki and Kazushi Ikeda. Adaptive Markov chain Monte Carlo for auxiliary variable method and its application to parallel tempering. *Neural Networks*, 43:33–40, 2013.

Yves Atchadé, Gersende Fort, Eric Moulines, and Pierre Priouret. Adaptive markov chain monte carlo: theory and methods. *Preprint*, 2009.

Yves Atchadé, Gersende Fort, et al. Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli*, 16(1):116–154, 2010.

Yves F Atchadé, Jeffrey S Rosenthal, et al. On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

Krishna B Athreya, Hani Doss, and Jayaram Sethuraman. A proof of convergence of the Markov chain simulation method. Technical report, DTIC Document, 1992.

Tim Austin. On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probability Surveys*, 5:80–145, 2008. doi: 10.1214/08-PS124. URL http://dx.doi.org/10.1214/08-PS124.

Yan Bai. An adaptive directional Metropolis-within-Gibbs algorithm. *Preprint*, 2009.

Yan Bai, Radu V Craiu, and Antonio F Di Narzo. Divide and conquer: a mixture-based approach to regional adaptation for MCMC. *Journal of Computational and Graphical Statistics*, 20(1):63–79, 2011.

LE Ballentine. Interpretations of probability and quantum theory. In *Foundations of probability and physics*, volume 1, pages 71–84. Quantum Probability White Noise Analysis, 2001.

Stefan Banach and Alfred Tarski. Sur la décomposition des ensembles de points en parties respectivement congruentes. *Fund. math*, 6(1):924, 1924.

Jørgen Bang-Jensen and Gregory Gutin. Theory, algorithms and applications. *Springer Monographs in Mathematics, Springer-Verlag London Ltd., London*, 2007.

Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL `http://www.sciencemag.org/content/286/5439/509.abstract`.

Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

Mr. Bayes and Mr Price. An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)*, pages 370–418, 1763.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2000. ISBN 9780471494645. URL `https://books.google.dk/books?id=IXyLQgAACAAJ`.

José Miguel Bernardo. *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting, September 6/10, 1983*, volume 2. Elsevier Science Ltd, 1985.

Jean Bertoin. Homogeneous fragmentation processes. *Probability Theory and Related Fields*, 121(3):301–318, 2001.

Jean Bertoin. *Random fragmentation and coagulation processes*, volume 102. Cambridge University Press, 2006.

G. Birkhoff and S.M. Lane. *A Survey of Modern Algebra*. AKP classics. Taylor & Francis, 1977. ISBN 9781568810683. URL `https://books.google.dk/books?id=FnP7sHxjt6gC`.

David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Charles Blundell and Yee Whye Teh. Bayesian Hierarchical Community Discovery. In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2013.

C. W Borchardt. Über eine Interpolationsformel für eine Art Symmetrischer Functionen und über Deren Anwendung. *Math. Abh. der Akademie der Wissenschaften zu Berlin*, pages 1–20, 1860.

P. J. Bowler. *Evolution: the history of an idea*. History of Sciencie. University of California Press, 1989. ISBN 978-0-520-06386-0. URL `http://books.google.ie/books?id=e2b5B0po8fwC`.

Anders Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, pages 929–953, 1999.

Stephen Brooks and Andrew Gelman. Some Issues for Monitoring Convergence of Iterative Simulations. *Computing Science and Statistics*, pages 30–36, 1998.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Brian Buck and Vincent A Macaulay. *Maximum entropy in action: a collection of expository essays*. Clarendon Press Oxford, 1991.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

Wray Buntine and Marcus Hutter. A Bayesian view of the Poisson-Dirichlet process. *arXiv preprint arXiv:1007.0296*, 2010.

Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468, 2000.

Penha Maria Cardoso Dias and Abner Shimony. A critique of Jaynes' maximum entropy principle. *Advances in Applied Mathematics*, 2(2):172–211, 1981.

Francois Caron. Bayesian nonparametric models for bipartite graphs. In *NIPS-Neural Information Processing Systems*, 2012.

Francois Caron and Emily B Fox. Bayesian nonparametric models of sparse and exchangeable random graphs. *arXiv preprint arXiv:1401.1137*, 2014.

Simon Carter, Marc Dymetman, and Guillaume Bouchard. Exact sampling and decoding in high-order hidden Markov models. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1125–1134. Association for Computational Linguistics, 2012.

George Casella and Edward I George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Ismaël Castillo. A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1-2):53–99, 2012.

Ariel Caticha. Lectures on Probability, Entropy, and Statistical Physics. *CoRR*, abs/0808.0012, 2008. URL `http://dblp.uni-trier.de/db/journals/corr/corr0808.html#abs-0808-0012`.

Ariel Caticha and Adom Giffin. Updating Probabilities. *CoRR*, abs/physics/0608185, 2006. URL `http://dblp.uni-trier.de/db/journals/corr/corr0608.html#abs-physics-0608185`.

A Cayley. A theorem on trees. *Quart. J. Math*, 23:376–378, 1889.

Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009.

David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

KS Chan. A note on the geometric ergodicity of a Markov chain. *Advances in Applied Probability*, pages 702–704, 1989.

G. Chartrand, L. Lesniak, and P. Zhang. *Graphs & Digraphs, Fifth Edition*. A Chapman & Hall book. Taylor & Francis, 2010. ISBN 9781439826270. URL `http://www.google.dk/books?id=K6-FvXRlKsQC`.

Sourav Chatterjee, Persi Diaconis, et al. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.

Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

Christophe Combet, Christophe Blanchet, Christophe Geourjon, and Gilbert Deleage. NPS@: network protein sequence analysis. *Trends in biochemical sciences*, 25(3):147–150, 2000.

D. Corfield and J. Williamson. *Foundations of Bayesianism*. Applied Logic Series. Springer, 2001. ISBN 9781402002236. URL `http://books.google.dk/books?id=74y__aTsckwC`.

A.A. Cournot. *Exposition de la théorie des chances et des probabilités*. L. Hachette, 1843. URL `https://books.google.com/books?id=_fk3AAAAMAAJ`.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

RT Cox. Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14:1–13, 1946.

R.T. Cox. *Algebra of Probable Inference*. Algebra of Probable Inference. Johns Hopkins University Press, 1961. ISBN 9780801869822. URL `http://books.google.dk/books?id=dcNpAUU6ACgC`.

Radu V Craiu, Jeffrey Rosenthal, and Chao Yang. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association*, 104(488):1454–1466, 2009.

Giulio D'Agostini. Bayesian reasoning versus conventional statistics in high energy physics. In *Maximum Entropy and Bayesian Methods Garching, Germany 1998*, pages 157–170. Springer, 1999.

George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8609–8613. IEEE, 2013.

Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Dover Publications, 2007.

Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, R Mena, I Prunster, and Matteo Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? 2013.

Pierre De Fermat and Etienne Pascal. 1654.

B. De Finetti. Funzione Caratteristica Di un Fenomeno Aleatorio. 6. Memorie, pages 251–299. Academia Nazionale del Linceo, 1931.

B. de Finetti. *Theory of probability: a critical introductory treatment*. Probability and Statistics Series. John Wiley & Sons Australia, Limited, 1974. ISBN 9780471201410. URL `https://books.google.dk/books?id=to0uAAAAIAAJ`.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

Pierre Simon de Laplace. *Théorie analytique des probabilités - 3ème édition*. Courcier, 1820.

Derek J de Solla Price. Is technology historically independent of science? A study in statistical historiography. *Technology and Culture*, pages 553–568, 1965.

Charo I Del Genio, Hyunju Kim, Zoltán Toroczkai, and Kevin E Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PloS one*, 5(4):e10012, 2010.

Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, pages 325–339, 1967.

Arnaud Doucet. *Sequential monte carlo methods*. Wiley Online Library, 2001.

JP Dougherty. Explaining statistical mechanics. *Studies in History and Philosophy of Science Part A*, 24(5):843–866, 1993.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Didier Dubois and Henri Prade. *Possibility theory*. Wiley Online Library, 1988.

Marijtje AJ Duijn, Tom AB Snijders, and Bonne JH Zijlstra. p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58 (2):234–254, 2004.

Maurice J Dupre and Frank J Tipler. The Cox Theorem: Unknowns And Plausible Value. *arXiv preprint math/0611795*, 2006.

Daniele Durante and David Dunson. {Bayesian Logistic Gaussian Process Models for Dynamic Networks}. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 194–201, 2014.

Rick Durrett. *Random graph dynamics*, volume 20. Cambridge university press, 2007.

Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

JR Ehrman, LD Fosdick, and DC Handscomb. Computation of Order Parameters in an Ising Lattice by the Monte Carlo Method. *Journal of Mathematical Physics*, 1:547–558, 1960.

Lloyd T Elliott and Yee Whye Teh. Scalable imputation of genetic data with a discrete fragmentation-coagulation process. In *NIPS*, pages 2861–2869, 2012.

R.L. Ellis. *On the Foundations of the Theory of Probabilities*. John William Parker, 1843. URL `https://books.google.com/books?id=in7IGwAACAAJ`.

P. Erdös and A. Rényi. On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959. URL `http://www.renyi.hu`.

Warren J Ewens. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112, 1972.

Dmitrii Konstantinovich Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Matematicheskikh Nauk*, 11(1):227–231, 1956.

Ruma Falk. When truisms clash: Coping with a counterintuitive problem concerning the notorious two-child family. *Thinking & Reasoning*, 17(4): 353–366, 2011. doi: 10.1080/13546783.2011.613690. URL `http://dx.doi.org/10.1080/13546783.2011.613690`.

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Thomas S Ferguson. Prior distributions on spaces of probability measures. *The annals of statistics*, pages 615–629, 1974.

Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: a survey. *arXiv preprint arXiv:1207.0246*, 2012.

Stephen E Fienberg. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21 (4):825–839, 2012.

Stephen E Fienberg et al. When did Bayesian inference become“ Bayesian”? *Bayesian analysis*, 1(1):1–40, 2006.

T.L. Fine. *Theories of probability: an examination of foundations*. Academic Press, 1973. URL `https://books.google.dk/books?id=U9lEAAAAIAAJ`.

L. D. Fosdick. . *Bull. Am. Phys. Soc*, 239, 1957.

A Stewart Fotheringham and Morton E O'Kelly. *Spatial interaction models: formulations and applications*. Kluwer Academic Dordrecht, 1989.

James R Foulds, Christopher DuBois, Arthur U Asuncion, Carter T Butts, and Padhraic Smyth. A dynamic relational infinite feature model for longitudinal social networks. In *International Conference on Artificial Intelligence and Statistics*, pages 287–295, 2011.

Ove Frank and David Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

Wenjie Fu, Le Song, and Eric P Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th annual international conference on machine learning*, pages 329–336. ACM, 2009.

M. Gardner. *The Scientific American book of mathematical puzzles & diversions*. Number v. 1 in The Scientific American Book of Mathematical Puzzles & Diversions. Simon and Schuster, 1959. URL `http://books.google.dk/books?id=KG5MAQAAIAAJ`.

Alan E Gelfand and Sujit K Sahu. On Markov chain Monte Carlo acceleration. *Journal of Computational and Graphical Statistics*, 3(3):261–276, 1994.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85 (410):398–409, 1990.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 11 1992. doi: 10.1214/ss/1177011136. URL `http://dx.doi.org/10.1214/ss/1177011136`.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Samuel J Gershman, Peter I Frazier, and David M Blei. Distance dependent infinite latent feature models. *arXiv preprint arXiv:1110.5454*, 2011.

John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department, 1991.

Subhashis Ghosal. The Dirichlet process, related priors and posterior asymptotics. In *Bayesian nonparametrics*, Camb. Ser. Stat. Probab. Math., pages 35–79. Cambridge Univ. Press, Cambridge, 2010.

Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001271. URL `http://dx.doi.org/10.1214/009053606000001271`.

W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):pp. 455–472, 1995. ISSN 00359254. URL `http://www.jstor.org/stable/2986138`.

Walter R Gilks. *Markov chain monte carlo*. Wiley Online Library, 2005.

Walter R Gilks, Gareth O Roberts, and Edward I George. Adaptive direction sampling. *The statistician*, pages 179–189, 1994.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2010.00765.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2010.00765.x`.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12): 7821–7826, 2002.

F.K. Glückstad, T. Herlau, M.N. Schmidt, and M. Morup. Unsupervised Knowledge Structuring: Application of Infinite Relational Models to the FCA Visualization. In *2013 International Conference on ignal-Image Technology Internet-Based Systems (SITIS)*, pages 233–240, Dec 2013a. doi: 10.1109/SITIS.2013.48.

Fumiko Kano Glückstad, Tue Herlau, Mikkel Nørgaard Schmidt, and Morten Mørup. *Analysis of Subjective Conceptualizations Towards Collective Conceptual Modelling*. Japanese Society for Artificial Intelligence, 2013b.

Fumiko Kano Glückstad, Tue Herlau, Mikkel Nørgaard Schmidt, Morten Mørup, Rafal Rzepka, and Kenji Araki. Analysis of Conceptualization Patterns across Groups of People. In *2013 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 349–354, 2013c.

Fumiko Kano Glückstad, Tue Herlau, Mikkel N Schmidt, and Morten Mørup. Cross-categorization of legal concepts across boundaries of legal systems: in consideration of inferential links. *Artificial Intelligence and Law*, pages 1–48, 2014.

Alexander Gnedin and Jim Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

M Grendár Jr and M Grendár. Maximum Probability and Maximum Entropy methods: bayesian interpretation. *arXiv preprint physics/0308005*, 2003.

Jim E. Griffin and Stephen G. Walker. Posterior Simulation of Normalized Random Measure Mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011. doi: 10.1198/jcgs.2010.08176. URL `http://amstat.tandfonline.com/doi/abs/10.1198/jcgs.2010.08176`.

TL Griffiths and Z Ghahramani. Infinite latent feature models and the indian buffet process. 2005.

Fan Guo, Steve Hanneke, Wenjie Fu, and Eric P Xing. Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pages 321–328. ACM, 2007.

Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

Bénédicte Haas, Grégory Miermont, Jim Pitman, Matthias Winkel, et al. Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. *The Annals of Probability*, 36(5):1790–1837, 2008.

Petr Hájek. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media, 1998.

Joseph Y Halpern. A counterexample to theorems of Cox and Fine. *J. Artif. Intell. Res.(JAIR)*, 10:67–85, 1999.

J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen's monographs on applied probability and statistics. Methuen, 1964. ISBN 9780416523409. URL `http://books.google.dk/books?id=Kk4OAAAAQAAJ`.

Mark S Handcock. Assessing Degeneracy in Statistical Models of Social Networks. 2003.

G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952. ISBN 9780521358804. URL http://books.google.dk/books?id=t1RCSP8YKt8C.

Michael Hardy. Scaled Boolean algebras. *Advances in Applied Mathematics*, 29(2):243–292, 2002. ISSN 0196-8858. doi: 10.1016/S0196-8858(02)00011-8. URL http://www.sciencedirect.com/science/article/pii/S0196885802000118.

J. Harris, J.L. Hirst, and M.J. Mossinghoff. *Combinatorics and Graph Theory*. Springer Undergraduate Texts in Mathematics and Technology. Springer, 2008. ISBN 9780387797106. URL http://books.google.dk/books?id=CxSoZcNymacC.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Creighton Heaukulani and Zoubin Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 275–283, 2013.

David Heckerman. An axiomatic framework for belief updates. In *UAI '86: Proceedings of the Second Annual Conference on Uncertainty in Artificial Intelligence, University of Pennsylvania, Philadelphia, PA, USA, August 8-10, 1986*, pages 11–22, 1986. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1764&proceeding_id=1002.

Tue Herlau, Morten Mørup, Mikkel N Schmidt, and Lars Kai Hansen. Detecting hierarchical structure in networks. In *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*, pages 1–6. IEEE, 2012a.

Tue Herlau, Morten Mørup, Mikkel N Schmidt, and Lars Kai Hansen. Modelling dense relational data. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012b.

Tue Herlau, Mikkel Schmidt, et al. Modeling temporal evolution and multiscale structure in networks. In *Proceedings of The 30th International Conference on Machine Learning*, pages 960–968, 2013.

Tue Herlau, Morten Mørup, Yee Whye Teh, and Mikkel N. Schmidt. Adaptive Reconfiguration Moves for Dirichlet Mixtures. page 26, May 2014a. URL http://arxiv.org/abs/1406.0071.

Tue Herlau, Mikkel N Schmidt, and Morten Mørup. Infinite-degree-corrected stochastic block model. *Physical Review E*, 90(3):032819, 2014b.

Tue Herlau, Morten Mørup, and Mikkel N. Schmidt. Bayesian Dropout. page 21, August 2015. URL `http://arxiv.org/abs/1508.02905`.

Edwin Hewitt and Leonard J Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, pages 470–501, 1955.

David M Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442): 585–595, 1998.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Nils Lid Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

Man-Wai Ho, Lancelot F James, and John W Lau. Coagulation fragmentation laws induced by general coagulations of two-parameter Poisson-Dirichlet processes. *arXiv preprint math/0601608*, 2006.

Qirong Ho, Ankur P Parikh, and Eric P Xing. A multiscale community blockmodel for network exploration. *Journal of the American Statistical Association*, 107(499):916–934, 2012.

Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*. MIT Press, 2007.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97 (460):1090–1098, 2002.

Jake M Hofman and Chris H Wiggins. Bayesian approach to network modularity. *Physical review letters*, 100(25):258701, 2008.

Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

D.N. Hoover. Relations on probability spaces and arrays of random variables, 1979.

Eric Horvitz, David Heckerman, and Curtis Langlotz. A Framework for Comparing Alternative Formalisms for Plausible Reasoning. In *AAAI*, pages 210–214, 1986.

Mark Huber. Efficient exact sampling from the Ising model using Swendsen-Wang. In *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 921–922. Society for Industrial and Applied Mathematics, 1999.

Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua B Tenenbaum. Dynamic Infinite Relational Model for Time-varying Relational Data Analysis. In *Advances in Neural Information Processing Systems*, pages 919–927, 2010.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.

Sonia Jain and Radford M Neal. A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004. doi: 10.1198/1061860043001. URL `http://amstat.tandfonline.com/doi/abs/10.1198/1061860043001`.

Lancelot F James. Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*, 2002.

Lancelot F James. A simple proof of the almost sure discreteness of a class of random measures. *Statistics & Probability Letters*, 65(4):363–368, 2003. ISSN 0167-7152. doi: 10.1016/j.spl.2003.08.005. URL `http://www.sciencedirect.com/science/article/pii/S0167715203002839`.

Lancelot F James. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Annals of statistics*, pages 1771–1799, 2005.

Lancelot F James. Poisson Dirichlet (α, θ)-Bridge Equations and Coagulation-Fragmentation Duality. *arXiv preprint arXiv:0908.4436*, 2009.

Lancelot F James. Stick-breaking PG (\ alpha,\ zeta)-Generalized Gamma Processes. *arXiv preprint arXiv:1308.6570*, 2013.

Lancelot F James, Antonio Lijoi, and Igor Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

Alejandro Jara, Emmanuel Lesaffre, Maria De Iorio, Fernando Quintana, et al. Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics*, 4(4):2126–2149, 2010.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957a.

Edwin T Jaynes. Information theory and statistical mechanics. II. *Physical review*, 108(2):171, 1957b.

Edwin T Jaynes. Where do we stand on maximum entropy. *The maximum entropy formalism*, pages 15–118, 1978.

Edwin T Jaynes. ET Jaynes: Papers on probability, statistics, and statistical physics. 1989.

Edwin T Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

R Johnson. Axiomatic characterization of the directed divergences and their linear combinations. *Information Theory, IEEE Transactions on*, 25(6):709–716, 1979.

RW Johnson and JE Shore. Comment on ''Consistent inference of probabilities for reproducible experiments''. *Physical review letters*, 55(3):336, 1985.

Dieter Jungnickel and Tilla Schade. *Graphs, networks and algorithms*, volume 5. Springer, 2005.

Olav Kallenberg. *Foundations of modern probability*. springer, 2002.

Olav Kallenberg. *Probabilistic symmetries and invariance principles*, volume 9. Springer, 2005.

Pl Kannappan. On Shannon's entropy, directed divergence and inaccuracy. *Probability Theory and Related Fields*, 22(2):95–100, 1972.

SN Karbelkar. On the axiomatic approach to the maximum entropy principle of inference. *Pramana*, 26(4):301–310, 1986.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Robert E Kass, Bradley P Carlin, Andrew Gelman, and Radford M Neal. Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

Aleksandr Iakovlevich Khinchin. *Mathematical foundations of information theory*, volume 434. Courier Dover Publications, 1957.

J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967. URL `http://projecteuclid.org/euclid.pjm/1102992601`.

J. F. C. Kingman. *Poisson Processes*. Oxford University Press, New York, NY, 1993.

JFC Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000029. URL `http://dx.doi.org/10.1214/009053606000000029`.

Kevin H Knuth and John Skilling. Foundations of inference. *Axioms*, 1(1): 38–73, 2012.

Teuvo Kohonen. Self-organization and associative memory. *Self-Organization and Associative Memory, 100 figs. XV, 312 pages.. Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences, volume 8*, 1, 1988.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Andrey Nikolaevich Kolmogorov. Foundations of probability. 1933.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, volume 1, page 4, 2012.

Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

Maciej Kurant, Athina Markopoulou, and Patrick Thiran. On the bias of bfs (breadth first search). In *Teletraffic Congress (ITC), 2010 22nd International*, pages 1–8. IEEE, 2010.

P.S. Laplace and A.I. Dale. *Pierre-Simon Laplace Philosophical Essay on Probabilities*. Sources in the History of Mathematics and Physical Sciences. Springer, 1995. ISBN 9780387943497. URL `https://books.google.com/books?id=vDZzuGcM4DUC`.

Krzysztof Łatuszyński, Gareth O Roberts, Jeffrey S Rosenthal, et al. Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability*, 23(1):66–98, 2013.

William H Lawton and Edward A Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3):617–633, 1971.

Wilhelm Lenz. Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. *Physikalische Zeitschrift*, 21(613-615):2, 1920.

Jure Leskovec. Nonparametric Multi-group Membership Model for Dynamic Networks. In *Neural Information Processing Systems*, 2013.

Faming Liang. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

Antonio Lijoi and Igor Prünster. Models beyond the Dirichlet process. *Bayesian nonparametrics*, 28:80, 2010.

Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291, 2005.

Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4): 769–786, 2007a.

Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007b.

Antonio Lijoi, Igor Prunster, and Stephen G Walker. Investigating nonparametric priors with Gibbs structure. *Statistica Sinica*, 18(4):1653, 2008.

Jun S Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

Jun S Liu. *Monte Carlo strategies in scientific computing*. springer, 2008.

Jun S Liu, Faming Liang, and Wing Hung Wong. The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.

Paul M Livingston. *Philosophical history and the problem of consciousness*. Cambridge University Press, 2004.

JR Lloyd, P Orbanz, Z Ghahramani, and D Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. *Advances in Neural Information Processing Systems*, 2013.

Peter McCullagh, Jim Pitman, Matthias Winkel, et al. Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002, 2008.

Aditya Krishna Menon and Charles Elkan. Predicting labels for dyadic data. *Data Mining and Knowledge Discovery*, 21(2):327–343, 2010.

Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer, 2011.

Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114. URL `http://scitation.aip.org/content/aip/journal/jcp/21/6/10.1063/1.1699114`.

Kurt T Miller, Thomas L Griffiths, and Michael I Jordan. Nonparametric latent feature models for link prediction. In *NIPS*, volume 9, pages 1276–1284, 2009.

Kurt Tadayuki Miller. *Bayesian nonparametric latent feature models*. PhD thesis, University of California, 2011.

Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

Andriy Mnih and Ruslan Salakhutdinov. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2007.

James Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, 69(2): 213–238, 2004.

M Morup, Mikkel N Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. In *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, pages 1–6. IEEE, 2011.

Morten Mørup and Mikkel N Schmidt. Bayesian community detection. *Neural computation*, 24(9):2434–2456, 2012.

D. J. Murdoch and P. J. Green. Exact Sampling from a Continuous State Space. *Scandinavian Journal of Statistics*, 25(3):483–502, 1998. ISSN 1467-9469. doi: 10.1111/1467-9469.00116. URL `http://dx.doi.org/10.1111/1467-9469.00116`.

Kevin P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, UBC, 2007.

Iain Murray, Zoubin Ghahramani, and David MacKay. MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.

Carlos Navarrete, Fernando A Quintana, and Peter Müller. Some issues in nonparametric Bayesian modeling using species sampling models. *Statistical Modelling*, 8(1):3–21, 2008.

Radford M Neal. Probabilistic inference using Markov chain Monte Carlo methods. 1993.

Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

MEJ Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, 2012.

Noam Nisan. *Algorithmic game theory*. Cambridge University Press, 2007.

E Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, Cambridge, 1984.

J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

Lars Onsager. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Phys. Rev.*, 65:117–149, Feb 1944. doi: 10.1103/PhysRev.65.117. URL http://link.aps.org/doi/10.1103/PhysRev.65.117.

Peter Orbanz. Lecture Notes on Bayesian Nonparametrics. *Journal of Mathematical Psychology*, 56:1–12, 2012.

Peter Orbanz and Daniel M Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *arXiv preprint arXiv:1312.7857*, 2013.

Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

John Paisley, David M. Blei, and Michael I. Jordan. Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference. 2014.

K Palla, D Knowles, and Z Ghahramani. An Infinite Latent Attribute Model for Network Data. In *International Conference on Machine Learning*, 2012.

Jeff B Paris. *The Uncertain Reasoner's Companion. Tracts in Theoretical Computer Science 39*. Cambridge University Press Cambridge, 1994.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

Oliver Penrose. Foundations of statistical mechanics. *Reports on Progress in Physics*, 42(12):1937–2006, 1979.

Mihael Perman, Jim Pitman, and Marc Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92 (1):21–39, 1992. ISSN 0178-8051. doi: 10.1007/BF01205234. URL http://dx.doi.org/10.1007/BF01205234.

Patrick O. Perry and Patrick J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849, 2013. ISSN 1467-9868. doi: 10.1111/rssb.12013. URL http://dx.doi.org/10.1111/rssb.12013.

Peter H Peskun. Optimum monte-carlo sampling using markov chains. *Biometrika*, 60(3):607–612, 1973.

J. Pitman. *Probability*. Springer Texts in Statistics. Springer, 1993. ISBN 9780387979748. URL http://books.google.dk/books?id=3ArvAAAAMAAJ.

J. Pitman and J. Picard. *Combinatorial Stochastic Processes:*. Combinatorial Stochastic Processes: École D'Été de Probabilités de Saint-Flour XXXII - 2002. Springer, 2006. ISBN 9783540309901. URL http://books.google.dk/books?id=6qFTR4PZE4AC.

Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902, 1999.

Jim Pitman. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability & Computing*, 11(5):501–514, 2002.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006.

S.D. Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier, 1837. URL `https://books.google.co.uk/books?id=s3YAAAAAMAAJ`.

G Polya. Mathematics and plausible reasoning. I. Induction and analogy in mathematics. II. Patterns of plausible inference. 1954.

William H Press, Brian P Flannery, Saul A Teukolsky, and William T Vetterling. Numerical recipes, 1990.

James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random structures and Algorithms*, 9(1-2):223–252, 1996.

Eugenio Regazzini, Antonio Lijoi, and Igor Prünster. Distributional results for means of normalized random measures with independent increments. *Annals of Statistics*, pages 560–585, 2003.

Hans Reichenbach. The theory of probability. An inquiry into the logical and mathematical foundations of the calculus of probability. 1950.

A. Rényi. *Wahrscheinlichkeitsrechnung: Mit einem Anhang über Informationstheorie*. Hochschulbücher für Mathematik. Deutscher Verlag der Wissenschaften, 1962. URL `http://books.google.dk/books?id=IA85AAAAIAAJ`.

Paul Ressel. De Finetti-type theorems: an analytical approach. *The Annals of Probability*, pages 898–922, 1985.

D. Revuz. *Markov Chains*. North-Holland mathematical library. North-Holland, 1975. ISBN 9780720424508. URL http://books.google.dk/books?id=YGtytgAACAAJ.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer, 1999.

Gareth O Roberts and Jeffrey S Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of applied probability*, pages 458–475, 2007.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

G.O. Roberts and O. Stramer. Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodology And Computing In Applied Probability*, 4(4):337–357, 2002. ISSN 1387-5841. doi: 10.1023/A:1023562417138. URL http://dx.doi.org/10.1023/A%3A1023562417138.

W.R. Roberts, I. Bywater, and F. Solmsen. *Aristotle: Rhetoric*. Modern library of the world's best books. Random House, 1954. URL https://books.google.com/books?id=az2NnQEACAAJ.

Garry Robins and Philippa Pattison. Random graph models for temporal processes in social networks*. *Journal of Mathematical Sociology*, 25(1):5–41, 2001.

Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430): 558–566, 1995a.

Jeffrey S Rosenthal. Rates of convergence for Gibbs sampling for variance component models. *The Annals of Statistics*, pages 740–761, 1995b.

Jeffrey S Rosenthal. A review of asymptotic convergence for general state space Markov chains. *Far East J. Theor. Stat*, 5(1):37–50, 2001.

Jeffrey S. Rosenthal. *A first look at rigorous probability theory*. World Scientific, Singapore [u.a.], 2. ed edition, 2006. ISBN 978-981-270371-2. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+529914859&sourceid=fbw_bibsonomy.

Daniel M Roy and Yee W Teh. The Mondrian Process. In *Advances in Neural Information Processing Systems*, pages 1377–1384, 2008.

Daniel M Roy, Charles Kemp, Vikash K Mansinghka, and Joshua B Tenenbaum. Learning annotated hierarchies from relational data. *Advances in neural information processing systems*, 19:1185, 2007.

Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 707. John Wiley & Sons, 2011.

B. Russell. *The Principles of Mathematics*. Number v. 1 in The Principles of Mathematics. University Press, 1903. URL `https://books.google.com/books?id=yN9LAAAAMAAJ`.

B. Russell. *History of Western Philosophy*. Routledge classics. Routledge, 1946. ISBN 9780415325059. URL `https://books.google.com/books?id=Ey94E3sOMA0C`.

Purnamrita Sarkar and Andrew W Moore. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40, 2005.

R. Scheaffer and L. Young. *Introduction to Probability and Its Applications*. Advanced series. Cengage Learning, 2009. ISBN 9780534386719. URL `http://books.google.dk/books?id=L95tVBSOgpUC`.

M. N. Schmidt, T. Herlau, and M. Mørup. Nonparametric Bayesian models of hierarchical structure in complex networks. *(Unpublished manuscript)*, sep 2012. URL `http://www2.imm.dtu.dk/pubdb/p.php?6522`.

Mikkel N Schmidt, Tue Herlau, and Morten Mørup. Nonparametric Bayesian models of hierarchical structure in complex networks. *arXiv preprint arXiv:1311.1033*, 2013.

Mikkel N. Schmidt, Tue Herlau, and Morten Mørup. Probabilistic structural hierarchical clustering of normal relational data. In *Cognitive Information Processing*, , 2014.

Amandine Schreck, Gersende Fort, and Eric Moulines. Adaptive Equi-Energy Sampler: Convergence and Illustration. *ACM Trans. Model. Comput. Simul.*, 23(1):5:1–5:27, January 2013. ISSN 1049-3301. doi: 10.1145/2414416.2414421. URL `http://doi.acm.org/10.1145/2414416.2414421`.

Ernst Schröder. XV. Vier combinatorische Probleme. *Zeitschrift für Mathematik und Physik*, 15:361–376, 1870.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991.

Glenn Shafer et al. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Madhusudana Shashanka, Bhiksha Raj, and Paris Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational intelligence and neuroscience*, 2008, 2008.

Abner Shimony. Comment on" Consistent inference of probabilities for reproducible experiments". *Physical review letters*, 55(9):1030, 1985a.

Abner Shimony. The status of the principle of maximum entropy. *Synthese*, 63 (1):35–53, 1985b.

John E. Shore and Rodney W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980. URL http://dblp.uni-trier.de/db/journals/tit/tit26.html#ShoreJ80.

Ajit P Singh and Geoffrey J Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases*, pages 358–373. Springer, 2008.

John Skilling. The axioms of maximum entropy. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, pages 173–187. Springer, 1988.

John Skilling. Classic maximum entropy. In *Maximum Entropy and Bayesian Methods*, pages 45–52. Springer, 1989.

Robert L Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

Ion Stoica, Robert Morris, David Karger, M Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *ACM SIGCOMM Computer Communication Review*, volume 31, pages 149–160. ACM, 2001.

Marshall H Stone. The theory of representation for Boolean algebras. *Transactions of the American Mathematical Society*, 40(1):37–111, 1936.

L. Strobel. *The Case for Faith*. Inspirio/Zondervan Miniature Editions. Running Press Book Publishers, 2004. ISBN 9780762421039. URL http://books.google.dk/books?id=czB7PwAACAAJ.

Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

Robert H Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical review letters*, 58(2):86–88, 1987.

Lei Tang and Huan Liu. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 817–826. ACM, 2009.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

Terence Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006.

Yee Whye Teh and Michael I Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics: Principles and Practice*, 28:158–207, 2010.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

Y Tikochinsky, NZ Tishby, and Raphael David Levine. Alternative approach to maximum-entropy inference. *Physical Review A*, 30(5):2638, 1984a.

Y Tikochinsky, NZ Tishby, and RD Levine. Consistent inference of probabilities for reproducible experiments. *Physical Review Letters*, 52(16):1357, 1984b.

Y Tikochinsky, NZ Tishby, and RD Levine. Tikochinsky, Tishby, and Levine respond. *Physical review letters*, 55(3):337, 1985.

M. Townsend. *Discrete mathematics: applied combinatorics and graph theory*. Benjamin/Cummings Pub. Co., 1987. ISBN 9780805393552. URL `http://books.google.dk/books?id=w-3uAAAAMAAJ`.

M. Tribus. *Rational descriptions, decisions, and designs*. Pergamon unified engineering series: engineering design section. Pergamon Press, 1969. URL `http://books.google.dk/books?id=RXuuAAAAIAAJ`.

Jos Uffink. Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Science. Part B. Studies in History and Philosophy of Modern Physics*, 26(3):223–261 (1996), 1995. ISSN 1355-2198. doi: 10.1016/1355-2198(95)00015-1. URL `http://dx.doi.org/10.1016/1355-2198(95)00015-1`.

Jos Uffink. The constraint rule of the maximum entropy principle. *Studies in History and Philosophy of Science. Part B. Studies in History and Philosophy of Modern Physics*, 27(1):47–79, 1996. ISSN 1355-2198. doi: 10.1016/1355-2198(95)00022-4. URL http://dx.doi.org/10.1016/1355-2198(95)00022-4.

J. P. Valleau and D. N. Card. Monte Carlo Estimation of the Free Energy by Multistage Sampling. *The Journal of Chemical Physics*, 57(12):5457–5462, 1972. doi: 10.1063/1.1678245. URL http://scitation.aip.org/content/aip/journal/jcp/57/12/10.1063/1.1678245.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008. ISSN 0090-5364. doi: 10.1214/009053607000000613. URL http://dx.doi.org/10.1214/009053607000000613.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1), 2001.

Bas C Van Fraassen. A problem for relative information minimizers in probability kinematics. *The British Journal for the Philosophy of Science*, 32(4):375–379, 1981.

Bas C Van Fraassen, RIG Hughes, and Gilbert Harman. A problem for relative information minimizers, continued. *British Journal for the Philosophy of Science*, pages 453–463, 1986.

Kevin S Van Horn. Constructing a logic of plausible inference: a guide to cox's theorem. *International Journal of Approximate Reasoning*, 34(1):3–24, 2003.

J. Venn. *The logic of chance: An essay on the foundations and province of the theory of probability, with especial reference to its application to moral and social science*. Macmillan, 1866. URL https://books.google.com/books?id=VAVVAAAAMAAJ.

Duy Quang Vu, Arthur U Asuncion, David R Hunter, and Padhraic Smyth. Continuous-Time Regression Models for Longitudinal Networks. In *NIPS*, pages 2492–2500, 2011.

Stephen Walker and Pietro Muliere. Beta-Stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics*, pages 1762–1780, 1997.

Sida Wang and Christopher Manning. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 118–126, 2013.

Stanley Wasserman and Carolyn Anderson. Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9(1):1–36, 1987.

Harrison C White, Scott A Boorman, and Ronald L Breiger. Social structure from multiple networks. I. Blockmodels of roles and positions. *American journal of sociology*, pages 730–780, 1976.

W. W. Wood and F. R. Parker. Monte Carlo Equation of State of Molecules Interacting with the Lennard-Jones Potential. I. A Supercritical Isotherm at about Twice the Critical Temperature. *The Journal of Chemical Physics*, 27 (3):720–733, 1957. doi: 10.1063/1.1743822. URL `http://scitation.aip.org/content/aip/journal/jcp/27/3/10.1063/1.1743822`.

William K Wootters. Statistical distance and Hilbert space. *Physical Review D*, 23(2):357, 1981.

Kevin S Xu and Alfred O Hero III. Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 201–210. Springer, 2013.

Zhao Xu, Volker Tresp, Kai Yu, and Hans-peter Kriegel. Infinite hidden relational models. In *In Proceedings of the 22nd International Conference on Uncertainity in Artificial Intelligence (UAI)*, 2006.

Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

Lotfi A Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *Systems, Man and Cybernetics, IEEE Transactions on*, (1):28–44, 1973.

Lotfi A Zadeh. The concept of a linguistic variable and its application to approximate reasoning—I. *Information sciences*, 8(3):199–249, 1975.

Lotfi A Zadeh. Discussion: Probability theory and fuzzy logic are complementary rather than competitive. *Technometrics*, 37(3):271–276, 1995.

Jörg Zimmermann and Armin B Cremers. *The quest for uncertainty*. Springer, 2011.